

LONDON
SCHOOL of
HYGIENE
& TROPICAL
MEDICINE



LSHTM Research Online

Aspeling-Jones, HMJ; (2019) Novel empirical and bioinformatic approaches to characterising *Plasmodium falciparum* antigens and their application to a merozoite-stage vaccine candidate. PhD thesis, London School of Hygiene & Tropical Medicine. DOI: <https://doi.org/10.17037/PUBS.04652254>

Downloaded from: <http://researchonline.lshtm.ac.uk/4652254/>

DOI: <https://doi.org/10.17037/PUBS.04652254>

Usage Guidelines:

Please refer to usage guidelines at <https://researchonline.lshtm.ac.uk/policies.html> or alternatively contact researchonline@lshtm.ac.uk.

Available under license: <http://creativecommons.org/licenses/by-nc-nd/2.5/>

<https://researchonline.lshtm.ac.uk>

LONDON
SCHOOL of
HYGIENE
& TROPICAL
MEDICINE



Novel empirical and bioinformatic approaches to characterising *Plasmodium falciparum* antigens and their application to a merozoite-stage vaccine candidate

Harvey Michael James Aspeling-Jones

Thesis submitted in accordance with the requirements for the degree of

Doctor of Philosophy

University of London

July 2017

Department of Pathogen Molecular Biology

Faculty of Infectious Tropical Diseases

LONDON SCHOOL OF HYGIENE & TROPICAL MEDICINE

Funded by Medical Research Council Vaccine Studentship

Research supervisor: Professor David Conway

Declaration of work

I Harvey Michael James Aspeling-Jones, confirm that the work presented in this thesis is my own.

Where information has been derived from other sources, I confirm that this has been indicated in the thesis.

Within chapters 2 and 4 blood samples were collected by directors, clinical staff and technicians of the Kintampo Municipal Hospital and the Kintampo Health Research Centre..

Within chapter 4 purification of MSP-1BLH was performed with assistance from Dr Sarah Tarr.

Preparation of B-cells for cell sorting was done with assistance from Dr Ofelia Diaz. Blood samples were collected from donors on the London School of Hygiene & Tropical Medicine anonymous blood donors' register by Carolynne Stanley. Ig gene amplification protocols were optimised by Miss Lindsay Serene under my supervision and with assistance from Mrs Lindsay Stewart.

Acknowledgements

I am indebted to my supervisor, Professor David Conway, for his continued support of my project and for the time he has invested in both developing this work and myself as a scientist. I am very grateful to Mrs Lindsay Stewart who was instrumental in both facilitating and guiding my work as well as encouraging me to work to a high standard. I would also like to thank Dr Craig Duffy, for providing tireless support especially with my computational work as well as being a great sounding board for ideas. I am very grateful to Sarah Tarr for her support, especially with protein purification and FACS work and her optimism. I owe many thanks to Dr Ofelia Diaz, who courageously volunteered to help with my cell sorting experiments whilst improving my Spanish and my mood. I would also like to thank Miss Henrietta Mensah Brown for fighting with me against stacks of ELISA plates with such grace. I am very grateful to Miss Lindsay Serene whose work optimising Ig gene PCR was much appreciated. I am lucky to have had Dr Sammuell Assefa, Dr Lee Murray, Miss Suzane Hocking and Dr Paul Divis as colleagues and each has provided useful insights during the course of my project. I owe much to Dr Kevin Tetteh who was very generous with his time, support and advice. I would also like to thank Dr Sam Alsford for his excellent work and support as research degrees coordinator. I am also grateful to Carolynne Stanley for accommodating my requests for blood samples so kindly.

I am indebted to Dr Gordon Awandare for his very generous support, without which my work in Ghana would not have been possible, and for his impeccable hospitality. I am also very thankful to Mr Rupert Delimini for working so hard to look after me in Kintampo on both a personal and professional level. I owe many thanks to Jones Opoku Mensah both for his tireless efforts to recruit donors and for welcoming me into his church. I am also thankful to the director and staff of Kintampo Health Research Centre for inviting me to work at their fantastic centre and making me feel so welcome.

I am very grateful to Professor James Beeson for hosting me on my placement at Burnett Institute and giving me the chance to work with his great team. I am also indebted to Dr Gaoqian Feng who patiently taught me how to perform invasion assays. My time at Burnett would not have been the same without Dr Vashti Irani, Dr Sarah Charnaud, Dr Brendan Elsworth and all the other students and staff who welcomed me into their community and made my time in Melbourne unforgettable.

I am indebted to the donors, both at the London School of Hygiene and Tropical Medicine and living and working in Kintampo, who gave up their precious time and blood to provide me with samples for the work presented here.

I must also thank my partner Georgina for her love and support through this journey, which would have been unbearable without her. I am also very grateful to my mother Jane and sister Izzy and to my whole family for continuing to put up with me and never letting me doubt they are there when I need them. Finally, a thank you to my friends and the players on my football team who have endured my poor friendship and even worse footballing skills to provide me with much needed distraction.

Abstract

A highly efficacious vaccine against the malaria parasite *Plasmodium falciparum* is needed. Repeat sequences are common in *P. falciparum* proteins and some are known immune targets. Short read sequence data are available for thousands of parasite isolates, but aligning and assembling repetitive sequences remains a challenge. A combined mapping and de novo assembly approach was developed to resolve highly complex and polymorphic allelic repeat sequences in the merozoite protein MSP-1. This approach gave an unbiased call of allele frequencies and full repeat sequence for a majority of clinical isolates tested. These data were used to design polyvalent hybrid sequences that would contain motifs from multiple alleles. Potential construct designs representing a greater spectrum of sequence diversity than that of previously designed polyvalent hybrid antigens were generated.

Assays of mechanisms of antibody mediated inhibition of parasite growth are needed to identify which antigen sequences are functional targets of immunity. Such assays are hard to standardise and would be benefitted by availability of human monoclonal antibody reagents. As an approach towards obtaining these, a technique was developed using a tetramerised *P. falciparum* MSP-1 recombinant antigen to isolate cognate B-cells from the blood of exposed Ghanaian adult donors. Despite lower than expected viability of cryopreserved samples, 82 memory antigen specific B-cells were successfully isolated by single cell flow sorting of lymphocytes from 16 donors. Complimentary DNA encoding both the heavy and light chain immunoglobulin variable regions was sequenced and analysed for two of these cells, revealing some distinct features. This is the first time a tetrameric antigen has been used to isolate human B-cells recognising a *P. falciparum* antigen, demonstrating their potential for use in the study of malarial immunity. The modest numbers of specific B-cells sorted from cryopreserved samples encourage the application of this approach to freshly obtained samples.

Table of contents

Declaration of work	2
Acknowledgements.....	3
Abstract.....	4
List of abbreviations.....	8
List of figures.....	10
List of tables	12
Chapter 1 - Introduction	13
1.1 Global burden of <i>P. falciparum</i> and need for efficacious vaccines.....	13
1.2 The lifecycle of <i>P. falciparum</i> offers multiple opportunities for a vaccine	16
1.3 Natural development of clinical immunity to <i>P. falciparum</i> is indicative of potential for effective vaccine development	19
1.4 Overcoming parasite diversity through designing blood-stage malaria vaccines	23
1.5 Repeat sequences are common features of <i>P. falciparum</i> antigens and have the potential to form the basis of multi-allelic vaccines.....	26
1.6 Challenges of extracting repeat sequence data with modern sequencing technologies.....	28
1.7 <i>P. falciparum</i> presents a wealth of potential vaccine antigens requiring characterisation and prioritisation	29
1.8 Correlates of protection from <i>P. falciparum</i> malaria are lacking	30
1.9 Human monoclonal reagents have the potential to inform design of vaccines aiming to block merozoite invasion of red blood cells.....	33
1.10 Antibodies against antigens on the merozoite surface can trigger secondary immune mechanisms leading to merozoite neutralisation	40
1.11 Antibodies target destruction of infected red blood cells.....	43
1.12 Recombinant human monoclonal antibodies enable the assaying of antibody mediated killing of <i>P. falciparum</i> merozoites and infected red blood cells induced by antibody responses to specific antigens.....	45
1.13 MSP-1 is processed before and during merozoite invasion and encodes tripeptide repeat sequences	46
1.14 Aims and objectives	49
Chapter 2 - Calling <i>msp1</i> block 2 alleles from short read data	60
2.1 Introduction	60
2.2 Materials and methods.....	70
2.2.1 Long read sequence data	70
2.2.2 Generation of synthetic reads from long read sequence data	70
2.2.3 Illumina paired-end short read sequence data.....	71
2.2.4 <i>De novo</i> assembly	71
2.2.5 Alignment of short reads	72

2.2.6 Data analysis	72
2.3 Results.....	74
2.3.1 Creation of long read sequence dataset for <i>msh1</i> block 2	74
2.3.2 <i>De novo</i> assembly optimised for reconstruction of <i>msh1</i> block 2 sequences	74
2.3.3 Optimised <i>de novo</i> assembly of <i>msh1</i> block 2 sequences results in bias towards short alleles	77
2.3.4 Creation of a reference library of <i>msh1</i> block 2 sequences allows for reads to be aligned	80
2.3.5 Alignment to library of <i>msh1</i> block 2 sequences enables unbiased calling of allelic type from synthetic short read data	84
2.3.6 Global distribution of <i>msh1</i> block 2 alleles as determined from short read data is similar to historical long read data	88
2.3.7 <i>De novo</i> assembly of library-aligned reads increases yield of sequences	96
2.3.8 Agreement between <i>de novo</i> assembly and alignment to the <i>msh1b2RefLib</i>	97
2.3.9 <i>Msh1</i> block 2 repeat lengths and structures determined by <i>de novo</i> assembly vary between Africa and Asia	100
2.4 Discussion.....	113
Chapter 3 - Obtaining a universal catalogue of MSP-1 block 2 epitope sequences from short read data	118
3.1 Introduction	118
3.2 Materials and methods.....	120
3.2.1 Data sources.....	120
3.2.2 Translation of aligned reads.....	121
3.2.3 Analysis of nonamers and design of minimal polyvalent antigens.....	121
3.3 Results.....	121
3.3.1 Short reads can be accurately translated based on alignment to a sequence library.....	121
3.3.2 Prevalence of nonamer epitopes varies by region	122
3.3.3 An algorithm for designing polyvalent hybrid antigens was designed and optimised	128
3.3.4. Design of region specific polyvalent antigens to incorporate range of epitopes	136
3.4 Discussion.....	143
Chapter 4 - Experimental approaches toward producing recombinant monoclonal antibodies against MSP-1.....	147
4.1 Introduction	147
4.2 Materials and methods.....	154
4.2.1 Polyvalent hybrid (PVH) MSP-1 antigens.....	154
4.2.2 Full length MSP-1 antigen	158
4.2.3 Tetramerisation of antigens.....	159
4.2.4 Sample collection	160

4.2.5 Preparation of B-cells.....	160
4.2.6 Amplification of Ig gene variable regions	161
4.2.7 Ig gene sequence analysis.....	161
4.3 Results.....	163
4.3.1 Cysteine residue successfully introduced to T-cell epitope of polyvalent hybrid MSP-1 antigen	163
4.3.2 Chemical biotinylation and tetramerisation of polyvalent hybrid antigen fails	166
4.3.3 Tetramerisation of biotinylated full-length MSP-1 was optimised.....	168
4.3.4 Isolation of MSP-1 specific memory B-cells	172
4.3.5 Ig gene variable regions sequenced for two antigen positive B-cells.....	177
4.4 Discussion.....	233
Chapter 5 - Discussion.....	238
5.1 The use of short read data for the analysis of population-wide variation in repeat sequences	238
5.2 The use of tetramerised antigens for the isolation of antigen-specific memory B-cells.....	240
5.3 Tools for the design and validation of vaccine antigens based on polymorphic repeat sequences	241
5.4 Concluding remarks	242
6. References	243
7. Appendices.....	279
7.1 <i>Msp1</i> block 2 genotyping studies.....	279
7.2 Long read sequences from GenBank in the long read dataset (LRD)	282
7.3 Python script for generating dummy reads	283
7.4 Sequences used in the <i>msp1b2RefLib</i>	285
7.5 Coverage of the <i>msp1b2RefLib</i> by reads from Pf3k data	287
7.6 Map showing location of sites of studies contributing to the Pf3k project.....	290
7.7 Python functions for translating aligned reads, obtaining and analysing nonamers	291
7.8 Python script for algorithm to design polyvalent hybrid antigens	295
7.9 FACS plots showing labelling of memory B-cells with MSP-1-SAPE antigen tetramers	297
7.10 List of additional data files	299
7.11 References	300

List of abbreviations

AARP-1 - asparagine and aspartate rich protein-1	LRD - long read dataset
AMA - apical merozoite antigen	LSA - liver-stage antigen
APC - allophycocyanin	mg - milligram
BAM - binary alignment/map	mL - millilitre
BASELINE - Bayesian estimation of Ag-driven selection	μL - microlitre
BB515 - BD Horizon Brilliant™ Blue 515	μm - micrometre
BCR- B-cell receptor	μM - micromolar
BioPVHA - biotinylated polyvalent hybrid antigen	MOI - multiplicity of infection
BLAST - basic local alignment search tool	MSPDBL - merozoite surface protein Duffy binding like proteins
BLH - biotinylation site linker histidine tag	MRIS - MR recombinant identifier sequence
bp - base pair	mRNA - messenger ribonucleic acid
BSA - bovine serum albumin	MSP - merozoite surface protein
C-terminus - carboxy-terminus	MVA - modified vaccinia virus Ankara
CD - cluster of differentiation	N-terminus - amino-terminus
cDNA - complementary deoxynucleic acid	ng - nanogram
CDR - complementarity determining region	nm - nanometre
ChAd63 - chimpanzee adenovirus 63	NTA - nitrilotriacetic acid
CS - Constant Spring	OLC - overlap layout consensus
CSP - circumsporozoite protein	<i>P. falciparum</i> - <i>Plasmodium falciparum</i>
CyRPA - cysteine-rich protective antigen	PAGE - polyacrylamide gel electrophoresis
DMSO - dimethyl sulfoxide	PBMC - peripheral blood mononuclear cells
DNA - deoxyribonucleic acid	PBS - phosphate buffered saline
DBG - De Bruijn graph	PCR - polymerase chain reaction
E. Africa - East Africa	PDB - Protein Data Bank
EBA - erythrocyte binding antigen	PEG - polyethylene glycol
E. Coli - <i>Escherichia coli</i>	PerCP - peridinin-chlorophyll-protein
EBV - Epstein Barr virus	<i>PfAPI</i> - <i>Plasmodium falciparum</i> annual parasite incidence
EG6PD - erythrocyte glucose-6-phosphate dehydrogenase	PfEMP-1 - <i>P. falciparum</i> erythrocyte membrane protein-1
EGF - epidermal growth factor-like	<i>PfPR</i> - <i>Plasmodium falciparum</i> parasite rate
EIR - entomological inoculation rate	PfRH - <i>Plasmodium falciparum</i> parasite rate
Fab - fragment antigen binding	pH - potential hydrogen
FBS - foetal bovine serum	PIPE - polymerase incomplete primer extension
Fc - fragment crystallisable	PNG - Papua New Guinea
FVO - <i>falciparum</i> Vietnam oak-knoll	PV - parasitophorous vacuole
FVS780 - fixable viability stain 780	PVH - polyvalent hybrid
GIA - growth inhibition assay	PVHA - polyvalent hybrid antigen
GLURP - glutamate rich protein	RAP - rhoptry-associated protein
GPI - glycosylphosphatidylinositol	RBC - red blood cell
GST - glutathione-S-transferase	RDT - rapid diagnostic test
Hb - Haemoglobin	RESA - ring-infected erythrocyte surface antigen
HIV - human immunodeficiency virus	RH - reticulocyte-binding protein homolog
Ig - immunoglobulin	RON - rhoptry neck protein
IgG - immunoglobulin G	R-PE - R-pyhoerythrin
IMGT - Immunogenetics	
kb - kilobase	

RT-PCR - reverse transcriptase polymerase chain reaction

S-antigen - soluble antigen

S. Asia - South Asia

S.E. Asia - South East Asia

SA - streptavidin

SAM - sequence alignment/map

SAPE - streptavidin-phycoerythrin

SD - standard deviation

SDS - sodium dodecyl sulphate

SERA - serine repeat antigen

sfp - seed forming propensity

SHM - somatic hypermutation

SNP - single nucleotide polymorphism

TigGER - tool for Ig genotype elucidation via rep-seq

TRAP - thrombospondin-related adhesive protein

UV - ultra violet

W. Africa - West Africa

List of figures

Figure 1.1 Map showing global prevalence of <i>P. falciparum</i>	15
Figure 1.2 Schematic representation of the life cycle of <i>P. falciparum</i> and merozoite cell structure .	19
Figure 1.3 Model predicted change in probability of severe malaria, clinical malaria and parasitaemia with age.....	22
Figure 1.4 Schematic representation of processing of MSP-1.....	48
Figure 2.1 Schematic representation of two main assembly algorithms	68
Figure 2.2 Schematic representation of how assembly algorithms represent repeat sequences	69
Figure 2.3 Flow chart showing how synthetic reads were generated from long read sequence data.	73
Figure 2.4. The effect of <i>k-mer</i> length on the fraction of <i>msp1</i> block 2 sequences assembled by Velvet.	76
Figure 2.5. Frequency distributions of length of block two sequence for assembled and unassembled sequences show bias towards assembly of shorter sequences.....	78
Figure 2.6 Probability of complete assembly of <i>msp1</i> block 2 is dependent on depth of coverage	79
Figure 2.7 Flow chart showing method for creaion of <i>msp-1</i> block 2 reference libraray.	82
Figure 2.8 Effect of the number of sequences in the reference library on the number of reads mapped	83
Figure 2.9. Distribution of coverage by allelic type after alignment of synthetic reads to <i>msp1b2RefLib</i>	86
Figure 2.10 Recombination of MAD20- and RO-33-like alleles during formation of MR recombinant allele creates unique sequence	87
Figure 2.11 Frequencies of alleles and mixed genotype infections vary by region.....	94
Figure 2.12 Allele frequencies across different study sites	96
Figure 2.13 Allele frequencies of K1-like and MAD20-like alleles show skews between continents .	109
Figure 2.14 RO-33-like alleles and their frequencies.....	110
Figure 2.15 MR alleles and their frequencies	111
Figure 2.16 Alignment of K1-like sequences suggests shared origin for most abundant alleles on each continent.....	112
Figure 3.1. Regional variation in frequencies of most common nonamers.....	128
Figure 3.2 Flow chart showing how polyvalent hybrid antigens were generated from short sequence reads	131
Figure 3.3 Optimisation of seed forming propensity for polyvalent hybrid antigen algorithm	132
Figure 3.4 <i>In silico</i> comparison of antigens designed by algorithm and by eye	134
Figure 3.5 Effect of length of antigen on coverage of sequences	135
Figure 3.6 Sequences of proposed polyvalent antigen designs.....	140
Figure 3.7 Nonamer peptide coverage of allele sequences in the tSRA dataset by proposed polyvalent hybrid antigens.....	142
Figure 4.1. Schematic representation of recombination at the IgH and Igκ/λ loci to produce heavy and light chain immunoglobulin transcripts	150
Figure 4.2 Schematic representation of BCR structure	152
Figure 4.3 Schematic representation of the introduction of a free sulfhydryl group into the polyvalent hybrid (PVH) MSP-1 antigen	157
Figure 4.4 Analysis of purified Bio-MSP1-BLH	159
Figure 4.5. Sequence alignment of PVH antigen F _{S26C} with PVH antigen F shows introduction of cysteine	164
Figure 4.6 Analysis of purified PVH antigens	165

Figure 4.7. Native protein gel showing no change in mass of biotinylated polyvalent hybrid antigen following incubation with streptavidin.....	167
Figure 4.8 Schematic representation of biotinylation of MSP-1 construct	170
Figure 4.9. Native PAGE showing change in mass of biotinylated MSP-1 following incubation with streptavidin	171
Figure 4.10. Gating strategy for isolation of MSP-1 specific B-cells	173
Figure 4.11. Comparison of MSP-1 positive memory B-cells between malaria exposed and naïve individuals	176
Figure 4.12. Deduced amino acid sequence of Ig variable regions from two MSP-1-specific memory B-cells	180
Figure 4.13 Alignment of EIMKB32-B11 IgH gene sequence with high affinity anti-MSP-119 Fab sequence.....	182
Figure 4.14. Alignment of IgH constant region containing rare variant present in three MSP-1-specific antibodies	232

List of tables

Table 1.1. Results of clinical trials of blood stage <i>P. falciparum</i> vaccines	51
Table 2.1 Distribution of msp1 block 2 allelic families by region from published studies	62
Table 2.2 Allele frequencies determined by alignment to a library of reference sequences.....	90
Table 2.3 Comparison of msp1 block 2 genotyping by alignment of short read data and published type-specific PCR.....	92
Table 2.4 Comparison of mixed genotype infections between Africa and Asia based on msp1 block 2 genotype	93
Table 2.5 Allelic frequencies of de novo assembled sequences	98
Table 2.6 Comparison of frequencies of allelic types called by alignment and de novo.....	99
Table 2.7 Number of unique peptide variants by region.....	101
Table 3.1. Coverage of nonamer epitopes in LRD sequences by antigens designed by algorithm using translated reads	133
Table 3.2 Comparison of coverage of MSP-1 block 2 sequences by novel designs for polyvalent hybrid antigens.....	139
Table 4.1 List of primer mixes used in amplification of Ig gene variable regions.....	162
Table 4.2 Cell counts for malaria exposed and naïve samples	175
Table 4.3 V(D)J gene usage for variable regions of two anti-MSP-1 antibodies.....	181
Table 4.4 Analysis of somatic hyper-mutation for evidence of antigen-driven selection	231

Chapter 1 - Introduction

1.1 Global burden of *P. falciparum* and need for efficacious vaccines

Human malaria is transmitted by the bite of mosquitoes infected with any of five *Plasmodium* species; *P. falciparum*, *P. vivax*, *P. malariae*, *P. ovale* and *P. knowlesi*. Malaria caused a global disease burden estimated at 214 million cases and 438 000 deaths in 2015 with two thirds of deaths occurring in children under five in sub-Saharan Africa as a result of *P. falciparum* infection (WHO, 2015). In the past 15 years there have been renewed international and regional initiatives to deploy new tools and drugs, namely insecticide-treated bed nets and artemisinin combination therapies, resulting in elimination in a few countries and substantial reductions in burden of disease in many other countries (figure 1.1). Targeted chemoprophylaxis has recently been employed in the Sahel region of Africa, and is proving to be successful at reducing the incidence of malaria amongst the population at highest risk (children and pregnant women). However, especially in hyper- and holo-endemic settings, new tools will be required to prevent deaths from malaria, which some estimates put as the leading cause of mortality in children under five worldwide (Elliott and Beeson, 2008). Furthermore, mosquito resistance to insecticides and the beginning of resistance to artemisinin in the parasite threaten to reverse the progress made over the last few decades (Phyo et al., 2012, Protopopoff et al., 2013).

Vaccines are highly efficacious public health tools, especially in tackling diseases of childhood, as they boost natural immunity and can provide life-long protection. Even before the molecular mechanisms were understood, vaccines were developed against smallpox and polio that have been used to eradicate and eliminate these once common diseases (Andre, 2003). Since the development of the first viral vaccines, global implementation of the MMR vaccine has led to dramatic reductions in the burden of mumps and measles, childhood diseases that were highly prevalent (Muller et al., 2007). Vaccines have also been used to reduce the burden of bacterial diseases including pertussis, diphtheria, meningococcus, pneumococcus, cholera, typhoid and *Haemophilus influenza* (Natalia et

al., 2009). Efficacious anti-parasite vaccines have been developed for use in veterinary medicine. These include vaccines against *Toxoplasma gondii* and *Emeria* parasites, apicomplexans related to *Plasmodium* species (although it is worth noting that these are whole-parasite vaccines) (Dalton and Mulcahy, 2001). However, despite several decades of research, a highly effective malaria vaccine is still not available (Matuschewski, 2017). RTS,S is the only licensed malaria vaccine, which has recently been recommended by the WHO for pilot implementation in sub-Saharan Africa (WHO, 2016). This vaccine contains epitopes from the circumsporozoite protein (CSP), which is expressed on the surface of the infective form of the parasite, and thus aims to boost immunity to initial infection and has been shown to be between 30% and 50% effective, although immunity wanes within a year of vaccination (RTS, 2015). Whilst this vaccine will might be useful in some situations, it is clear that an improved vaccine will be required in order to provide long lasting protection from disease. Whilst the RTS,S vaccine reduces the incidence of infection with *P. falciparum*, in those who are still infected it does not boost immune responses to the blood-stage form of the parasite, which is responsible for all disease symptoms. It is therefore important that *P. falciparum* blood-stage antigens are developed as vaccine targets, which could be co-formulated with RTS,S or used in other multi-stage formulations.

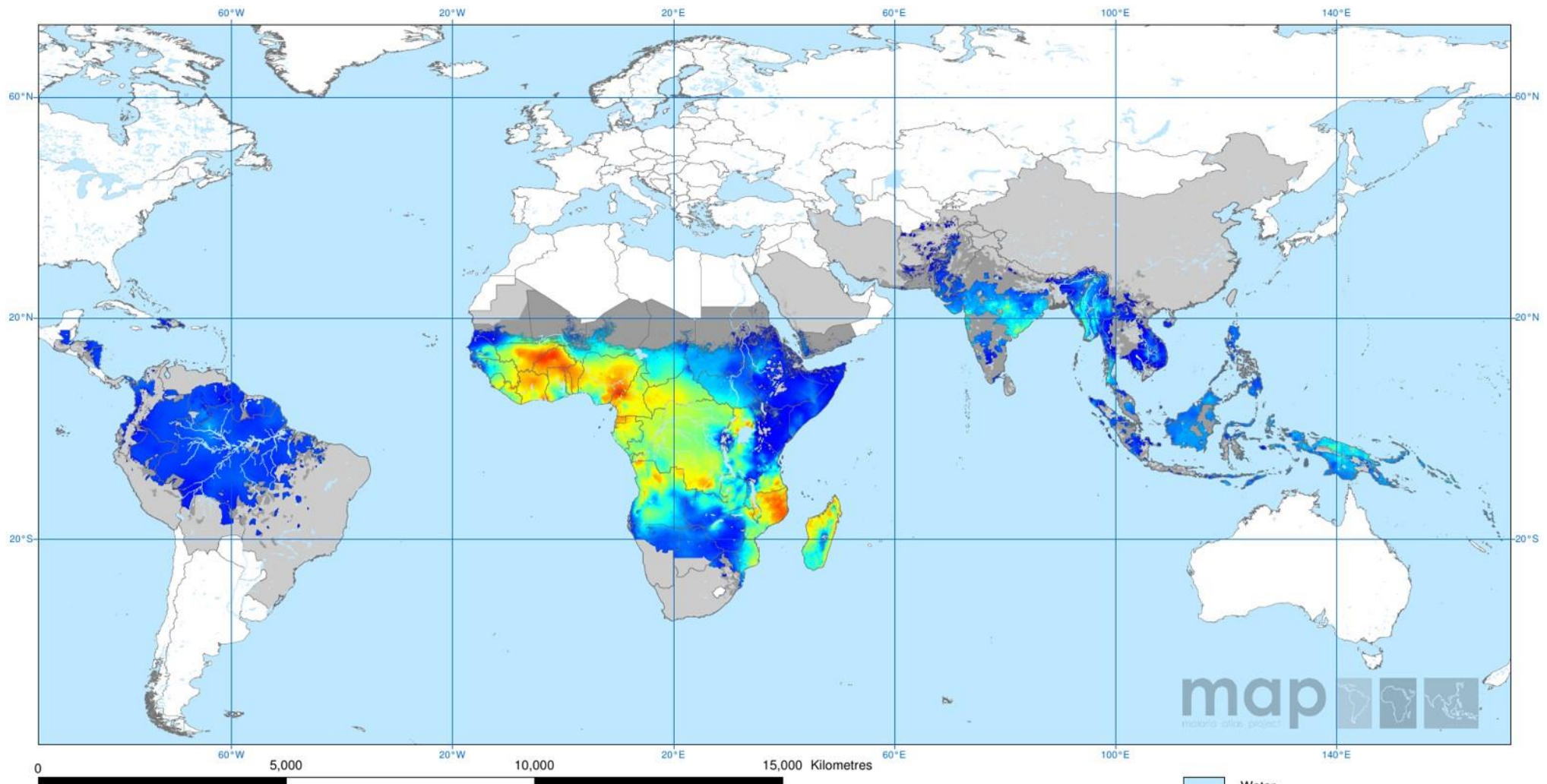


Figure 1.1 Map showing global prevalence of *P. falciparum*. The map is coloured according to prevalence of *P. falciparum* with a continuum from dark blue (0%) to red (70%) indicating the age-standardised annual mean percentage of children aged 2-10 years predicted to be positive for *P. falciparum* parasites ($PfPR_{2-10}$) in 2010 based on surveys of parasite prevalence conducted between January 1985 and June 2010. Countries with unstable transmission, defined as an annual incidence of *P. falciparum* ($PfAPI$) of less than 0.1 per 1000 people in 2010, or no transmission, defined as $PfAPI$ of zero in 2010, are coloured dark and light grey, respectively. Countries not contributing data are coloured white. Figure adapted from Gething *et al* 2011.

1.2 The lifecycle of *P. falciparum* offers multiple opportunities for a vaccine

All *Plasmodium* species infecting humans share a common lifecycle. The human infective form of the parasite, the sporozoite, is injected into the skin from the salivary glands of an infected female anopheline mosquito. Sporozoites then travel to the liver where they invade a hepatocyte and undergo schizogony to produce tens of thousands of merozoites which are then released into the blood. Once in the blood, parasites invade red blood cells (RBCs) and develop within a parasitophorous vacuole (PV), which, in the case of *P. falciparum*, takes around 48 hours and involves the transition through ring, trophozoite and schizont morphological stages. During this process, the parasite exports proteins to modulate the host cell membrane and promote adherence to endothelial cells (cytoadherence), presumably to avoid clearance in the spleen. At the end of intraerythrocytic cycle *P. falciparum* undergoes schizogony to produce an average of 16 daughter merozoites (Bannister and Mitchell, 2003), which are released following the breakdown of both the erythrocyte membrane and PV membrane and rapidly re-invade uninfected RBCs and establish another cycle of asexual reproduction. This haploid stage of the parasite life-cycle is the sole cause of pathology and also generates gametocytes, which result from the commitment of a small number (<10%) of schizonts to sexual development, known as gametocytogenesis, which produces mature male and female gametocytes after 10-12 days (Josling and Llinas, 2015). Upon uptake by a feeding mosquito gametocytes undergo rapid DNA replication and development to form gametes which fuse to form diploid zygotes that develop into motile ookinetes that infect the mosquito midgut wall. Ookinetes then develop into oocysts inside of which meiosis produces hundreds to thousands of haploid sporozoites (Stone et al., 2013) that are released on rupture of the oocyst and migrate to the salivary glands of the infected mosquito, ready to infect another human host (figure 1.2).

Vaccination against malaria has the potential to interrupt any stage of the parasite life-cycle in the human host and also to prevent transmission (figure 1.2). Pre-erythrocytic vaccines, including RTS,S, have been designed to elicit immune responses against sporozoites or liver-stage parasites to

prevent development of a blood stage infection. Blood-stage vaccines either aim to inhibit parasite replication through targeting merozoites or to block cytoadherence of infected RBCs (Richards and Beeson, 2009, Richie and Saul, 2002, Good et al., 1998). Transmission blocking vaccines aim to raise antibodies against gametocyte antigens that resulting in prevention of infection of the mosquito.

The merozoite is of particular importance to the design of blood-stage vaccines as, despite being short-lived, it is the only extracellular stage in which the parasite is fully exposed to the host immune system. The merozoite is a small (1.5 μm long) tear-drop shaped cell that has two major specialised compartments at the apical end, termed micronemes and rhoptries (figure 1.2) that contain proteins essential for invasion of the RBCs. The surface of the merozoite is decorated with a coat of largely peripheral and glycosylphosphatidylinositol (GPI)-anchored merozoite surface proteins (MSPs) which are proposed to mediate initial contact with the RBC membrane. This first step in the invasion process is followed by the re-orientation of the merozoite such that the apical end is in contact with the RBC. Invasion ligands are then released from the apical organelles (micronemes and rhoptries) which bind receptors on the RBC surface and commit the merozoite to invasion, a process in which a tight junction is formed with the host cell membrane and an actin-myosin motor drives the merozoite into an invagination of the RBC surface membrane that will become the PV (Cowman et al., 2012). Proteolytic processing of many MSPs and invasion ligands is essential for the merozoite to invade the host cell, and results in the shedding of many of these proteins during the invasion process (Beeson et al., 2016).

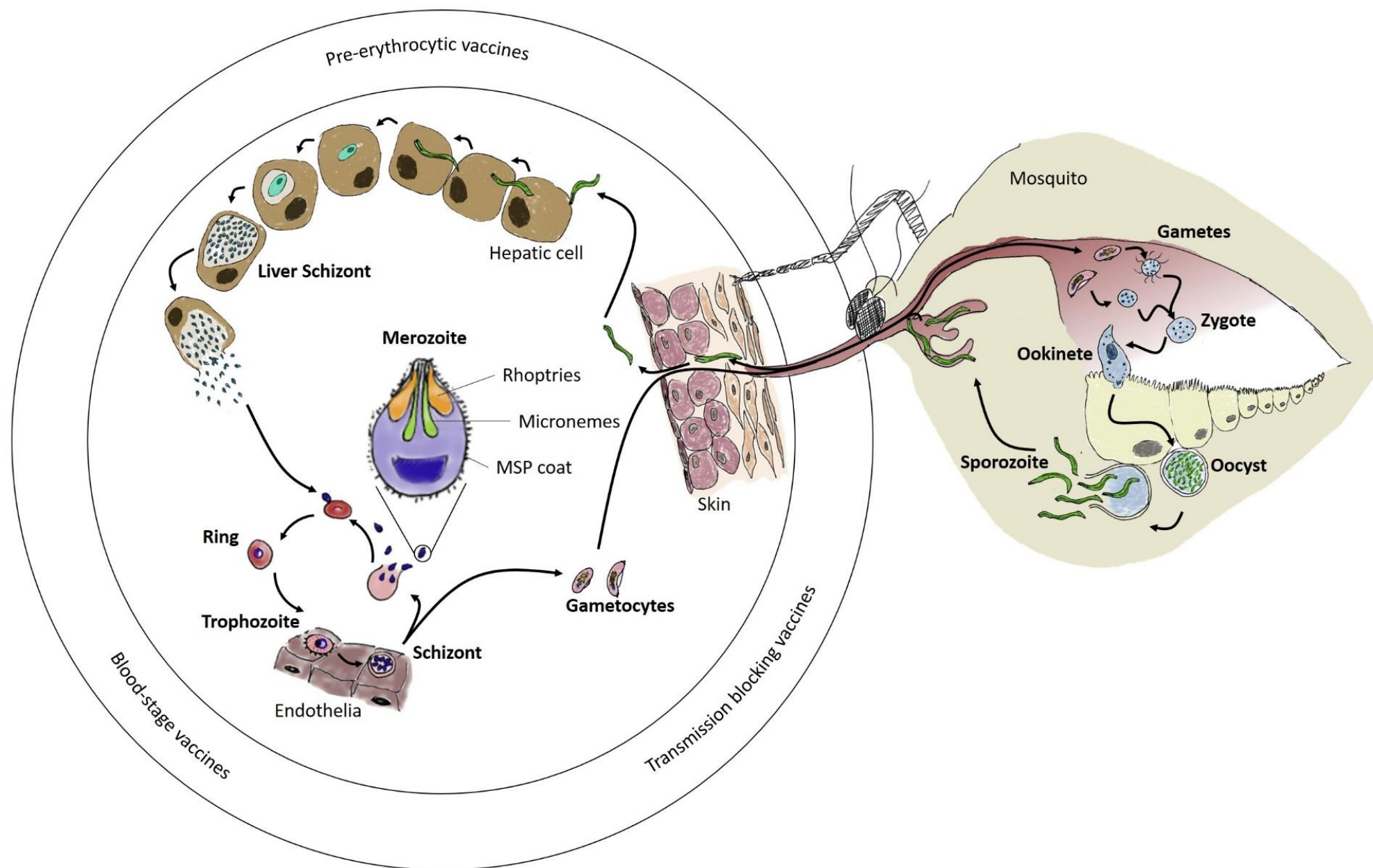


Figure 1.2 Schematic representation of the life cycle of *P. falciparum* and merozoite cell structure.

Sporozoites travel from the salivary glands of the infected female mosquitoes via the skin to the liver where they invade hepatocytes. *P. falciparum* develops within the hepatocyte, generating thousands of merozoites which are then released into the blood where they invade RBCs to establish a blood-stage infection. The parasite undergoes multiple cycles of asexual replication in the blood in which schizogony produces merozoites which invade fresh RBCs on rupture of the parent cell. The majority of schizonts produce merozoites that invade RBCs and develop into ring stage parasites which then develop into larger, amorphous trophozoites which modulate the RBC membrane to effect cytoadherence to the endothelia of microvasculature. A small minority of schizonts will commit to sexual development, producing merozoites that will invade RBCs to form rings stages but then go on to produce mature gametocytes which, following uptake by a feeding mosquito, will undergo sexual reproduction to produce mosquito invasive ookinets that go on to produce sporozoites that can establish infection of the next human host. The three vaccination strategies are shown at the edge of the life-cycle. Authors own representation of malaria life cycle.

1.3 Natural development of clinical immunity to *P. falciparum* is indicative of potential for effective vaccine development

In malarial, sterile immunity is the development of immune responses that prevent the parasite from establishing an infection. In contrast, clinical immunity is a state in which the host can be infected with parasites but does not develop disease symptoms. Although sterile immunity to *P. falciparum* is rarely, if ever seen, partially protective acquired immunity can cause a reduction of parasite numbers in the blood by four or five orders of magnitude and prevent disease (Druihe and Perignon, 1997); figure 1.3). Therefore it seems pertinent to pursue a vaccine that can induce clinical immunity to *P. falciparum*, which means developing a vaccine against the blood-stage of the parasite lifecycle.

In the 1960s, treatment with immunoglobulin gamma (IgG) from clinically immune adults was shown to reduce parasite burden and cure symptoms in children suffering from malaria, demonstrating that this class of antibodies are the key component of clinical immunity (Cohen et al., 1961, Edozien et al., 1962, McGregor, 1964). These studies were replicated later with non-immune Thai adults as recipients (Bouharoun-Tayoun et al., 1990, Sabchareon et al., 1991). These findings suggest that vaccination with blood stage antigens could result in clinical immunity.

The fact that clinical immunity takes a long time to develop and has been reported to wane in the absence of repeated exposure has been used to argue that there may be a defect in the formation of immunological memory (Langhorne et al., 2008). The detection of asymptomatic infections in low-transmission regions suggests, however, that clinical immunity is maintained despite sporadic exposure (Luxemburger et al., 1997, Alves et al., 2005, Branch et al., 2005, Cucunuba et al., 2008, Fugikaha et al., 2007, Roper et al., 2000, Roshanravan et al., 2003). Whilst it could be argued that clinical immunity to *P. falciparum* in low-transmission settings may not equate to clinical immunity in Africa due to the reduced parasite genetic diversity, similar observations were made in low transmission regions of Africa (Kleinschmidt and Sharp, 2001, Ouldabdallahi Moukah et al., 2016) where parasite genetic diversity is maintained by gene flow from neighbouring, high-endemic regions (Duffy et al., 2017). Studies comparing travellers who contract *P. falciparum* malaria found that, whilst both naïve and previously exposed individuals suffer morbidity, previously exposed patients cleared parasites faster, suffered milder disease symptoms and were less likely to die (Jelinek et al., 2002, Matteelli et al., 1999), although this finding was not replicated in a smaller study (Jennings et al., 2006). Malaria was eliminated from Madagascar in 1960 and was absent from the country until 1987 which marked the start of repeated outbreaks (Lepers et al., 1988, Romi et al., 2002). A lower incidence of malarial fever was observed in people over 40, although the rate of infection in this group was the same as the younger population, providing evidence that a degree of clinical immunity persists in the absence of repeated exposure. Multiple studies have observed that, in the absence of infection, antibodies against *P. falciparum* blood stage antigens decrease but are quickly boosted on re-infection, indicating that whilst antibody responses in the absence of antigen may be short-lived, memory B-cells persist (Cavanagh et al., 1998, Fruh et al., 1991, Taylor et al., 1998, Perraut et al., 2000, Jouin et al., 2001). Infection in a human challenge model has demonstrated the boosting of vaccine induced antibodies against two *P. falciparum* blood stage antigens by low-level parasitaemia (Elias et al., 2014).

The fact that the host immune response to *P. falciparum*, as with many parasitic organisms, appears to establish a balance in which low-level infection is tolerated (Bruce-Chwatt, 1963, Artavanis-Tsakonas et al., 2003) suggests that a blood-stage vaccine against *P. falciparum* may not be able to eliminate the disease, from either individuals or the population but, by pre-empting naturally acquired clinical immunity, such a vaccine would result in a massive reduction in the burden of disease.

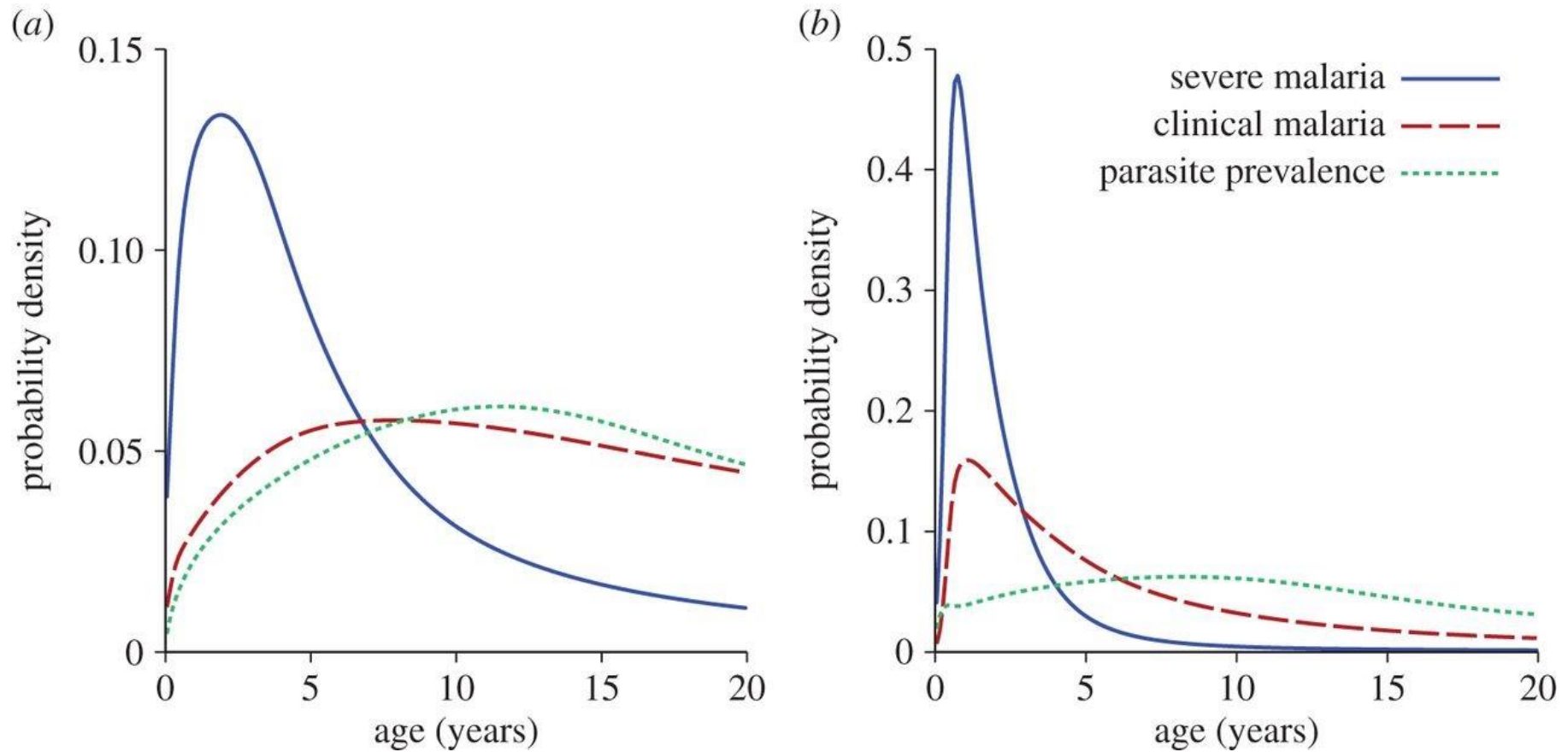


Figure 1.3 Model predicted change in probability of severe malaria, clinical malaria and parasitaemia with age. Griffin *et al* (2015) fitted a model of malaria transmission with the incidence of severe malaria by age obtained from hospital admission data collected in Africa. The change with age in (a) low (annual entomological inoculation rate (EIR) of 2) and (b) high (annual EIR of 50) transmission settings of the probability of severe malaria (blue), clinical malaria (red) parasite prevalence (green) as predicted by the model are shown. Increased exposure in high transmission settings results in a higher probability of disease but clinical immunity to disease develops faster. In both high and low transmission settings immunity to parasites develops slower than immunity to disease. Figure adapted from (Griffin *et al.*, 2015).

1.4 Overcoming parasite diversity through designing blood-stage malaria vaccines

Whilst clinically immune individuals control parasitaemia, they very rarely develop responses that inhibit mosquito infection. The infected mosquito may then pass the parasite onto a naïve host for whom the infection may prove fatal. The development of an immune response that controls the replication of parasites in the blood whilst maintaining infectivity is evidence of parasite-host co-evolution over many thousands of years. The establishment of this balance in naïve individuals is unlikely to be a straightforward task achieved by a single-antigen, mono-allelic vaccine.

The use controlled of *P. falciparum* infection in the treatment of neurosyphilis provided the first direct evidence of development of clinical immunity as patients exhibited reduced symptoms on secondary infection (Collins and Jeffery, 1999b). However, analysis of the records of controlled *P. falciparum* infection showed that clinical immunity developed primarily against the homologous parasite strain (Collins and Jeffery, 1999a). This finding indicates one of the major challenges for malaria vaccine design; the high degree of antigenic polymorphism and variation in antigen expression that allows the parasite to evade immunological memory. Indeed, the repeated episodes of malaria experienced as clinical immunity develops can be explained by strain specific clinical immunity (Bull et al., 1998, Roper et al., 1996).

Malaria vaccine antigens are generally polymorphic in natural parasite populations with some blood-stage antigens exhibiting very high levels of polymorphism (Takala and Plowe, 2009, Ouattara et al., 2015, Barry et al., 2009, Volkman et al., 2002, Polley and Conway, 2001, Miller et al., 1993, Smythe et al., 1990). The high degree of polymorphism is likely maintained by balancing selection exerted by the host immune system (Amambua-Ngwa et al., 2012b, Weedall and Conway, 2010). That these antigens appear to be under selection from the immune system provides evidence that they could potentially form the basis of a vaccine designed to offer clinical immunity. However, antigenic polymorphism poses a problem for vaccine design, as the presence of multiple alleles presents the opportunity for vaccine escape. Indeed, two trials of blood stage vaccines that showed little or no

vaccine efficacy did find strain-specific protective effects (Thera et al., 2011, Genton et al., 2002), which have been validated by *in vitro* analysis (Dutta et al., 2007, Ouattara et al., 2013, Stubbs et al., 2011).

Vaccines containing multiple allelic variants of antigens have been successfully used to immunise against both viral and bacterial diseases (Salk, 1953, Black et al., 2000). In the case of the influenza vaccine, antigenic variants are selected annually in response to predictions of the most prevalent strains (Ampofo et al., 2012). Vaccines containing multiple alleles have been proposed for apical merozoite antigen-1 (AMA-1), a region of merozoite surface protein-1 (MSP-1) and MSP-2, leading *P. falciparum* blood-stage vaccine candidates (Cowan et al., 2011, Krishnarjuna et al., 2016, Remarque et al., 2008, Tetteh and Conway, 2011, McCarthy et al., 2011). The only such vaccine to be subjected to a phase II trial contained two alleles of AMA-1 and was found to have no impact on parasite densities of vaccinated children (Sagara et al., 2009). However, AMA-1 contains many polymorphic, discontinuous epitopes (Duan et al., 2008, Escalante et al., 2001, Healer et al., 2002, Marshall et al., 1996, Polley and Conway, 2001) which may facilitate vaccine escape even when two alleles are present in the vaccine. Vaccines designed to elicit immune responses against a broader range of AMA-1 epitopes are currently being developed (Faber et al., 2016).

Whilst many cohort studies have found correlations between antibodies against blood stage antigens and protection from disease, results from separate studies are sometimes contradictory and a single leading blood-stage antigen is yet to emerge (reviewed in (Fowkes et al., 2010)). Studies that examined the response to panels of blood stage antigens have demonstrated that antibody responses to particular combinations of these antigens are most highly associated with protection (Osier et al., 2014b, Osier et al., 2008, Richards et al., 2013). This suggests that clinical immunity depends on antibody responses to a range of blood-stage antigens. For this reason, an efficacious blood-stage antigen is likely to require the combination of not just multiple allelic variants of an antigen, but multiple antigens as well. So far, only a few blood-stage vaccines comprising either fusion proteins containing domains of two antigens (Theisen et al., 2004, Hu et al., 2008) or mixtures

of two (Chitnis et al., 2015) or three (Saul et al., 1999) antigens have been entered into clinical trials. Results of phase II trials with two of these vaccines, whilst not showing high vaccine efficacy, encourage further development of multi-antigen vaccines (Sirima et al., 2016, Genton et al., 2002). Inoculation with irradiated sporozoites has shown that presentation of multiple pre-erythrocytic antigens leads to short-lived protection from infection (Seder et al., 2013). Infection of four subjects with low numbers of live blood-stage parasites followed by drug cure resulted in sterile immunity towards the homologous parasite strain in three subjects and reduced parasite growth in the fourth, indicating the potential of attenuated parasite vaccines (Pombo et al., 2002). However, the widespread use of whole organism parasite vaccines in endemic settings is prevented by technical barriers and cost of manufacture. It is therefore desirable to design a sub-unit or viral-vectored vaccine containing antigens that present a range of epitopes representing a selected combination of sequences of relevant blood-stage antigens.

At the time of writing, 54 trials of vaccines containing one or more blood-stage antigen have been published (table 1.1). Around half (28) of these trials were conducted with a vaccine containing only a single blood-stage antigen. Seven trials assessed efficacy of vaccines containing a single blood-stage antigen, with three finding evidence of protection (Ogutu et al., 2009, Sheehy et al., 2012, Sirima et al., 2011, Thompson et al., 2008, Thera et al., 2011, Laurens et al., 2017, Palacpac et al., 2013, Yagi et al., 2016), although one of these studies was underpowered (Sirima et al., 2011) and one was not replicated (Laurens et al., 2017). Excluding trials of SPf66, which are the subject of a Cochrane review (Graves and Gelband, 2006), six trials assessed vaccine efficacy for vaccines containing multiple antigens or multiple antigen alleles and half of these found some degree of protection (Lawrence et al., 2000, Sheehy et al., 2012, Sagara et al., 2009, Ockenhouse et al., 1998, Genton et al., 2002, Sirima et al., 2016). Although none showed efficacy greater than RTS,S, the results of these trials demonstrate the potential for a blood stage *P. falciparum* vaccine (Beeson et al., 2016).

Whilst 13 trials have focused on vaccines comprising multiple different antigens, just three have been conducted for vaccines containing multiple antigen alleles and only one of these had multiple alleles for multiple antigens. All of these trials demonstrated the potential to raise antibodies against both alleles present in the vaccine (Sagara et al., 2009, McCarthy et al., 2011, Ellis et al., 2012).

Whilst the only such trial to evaluate vaccine efficacy did not demonstrate protection (Sagara et al., 2009), this should not discourage the continued development of vaccines design to raise antibodies against multiple strains of polymorphic antigens.

Viral vectored vaccines not only offer a platform for the presentation of multiple antigens but can also be designed to elicit strong immune responses that can be polarised toward a cellular (T helper 1) or a antibody (T helper 2) response (Rollier et al., 2011). Seven antigens, from sporozoites, liver stages, blood-stages and gametocytes were included in a single vaccinia virus vectored vaccine (Tine et al., 1996), which was shown to cause a slight delay in the onset of parasitaemia in a human challenge model (Ockenhouse et al., 1998). However, no subjects sero-converted for all seven vaccines and immunogenicity was variable (Ockenhouse et al., 1998). Whilst more recent studies with viral vectored vaccines have not demonstrated efficacy in the same human challenge model (Sheehy et al., 2012), they do however suggest that vaccination with multiple viruses expressing different antigens may prove to be a better method for ensuring immune responses to all vaccine antigens and demonstrate the ability to elicit strong humoral and cellular immune responses (Biswas et al., 2014, Elias et al., 2014). Whilst issues remain regarding pre-existing immunity to the viral vector itself, viral vectored vaccines promise to provide a platform for the delivery of multiple alleles of multiple *P. falciparum* antigens (Rollier et al., 2011).

1.5 Repeat sequences are common features of *P. falciparum* antigens and have the potential to form the basis of multi-allelic vaccines

Tandem repeat sequences are DNA sequences that contain a sequence of bases that is repeated two or more times with no other sequences in between. These repeats can be perfect, comprising

exactly the same sequence repeated each time, or imperfect, where the repeat sequence varies.

Repeat sequences in protein coding regions of DNA will encode repeat sequence in the protein. The repeat sequences of *P. falciparum* proteins offer an attractive platform for the design of antigens presenting a range of common, linear B-cell epitopes that would allow the combination of multiple alleles of multiple blood-stage antigens required to elicit clinical immunity in naïve individuals.

Previous work demonstrated the abundance of repeat sequences in *P. falciparum* genes, being found in 56% (3058) of known protein coding genes and constituting 9.5% of coding sequence (Aspeling-Jones, 2013). The overwhelming majority of peptides encoded by these sequences are predicted to be intrinsically disordered (Feng et al., 2006). This lack of secondary and tertiary protein structure results in the presentation of linear B-cell epitopes, predicted to be enriched in intrinsically disordered protein domains (Guy et al., 2015). Moreover, repetitive sequences are likely to increase the affinity of antibody binding due to the concentration of binding sites.

Antibodies against the repeat sequences of the pre-erythrocytic antigen in RTS,S have been shown to be key for the protection offered by this vaccine (Kester et al., 2009, Foquet et al., 2014, Olotu et al., 2011). Antibodies against blood-stage antigens containing extensive repeat sequence have been shown correlate with protection from malarial disease (Taylor et al., 1995, Osier et al., 2010, Polley et al., 2007) and, in the case of MSP-1, antibodies against the repeat sequence show a stronger association with protection than those against the neighbouring non-repetitive sequences (Polley et al., 2003b). Many of the polymorphic sequences found in antigens are intrinsically disordered repeat (Anders et al., 1988, Anders et al., 1993, Tetteh et al., 2009, Cowman et al., 1985).

The amino acid moieties of a protein molecule that are recognised by an antibody are known as the antibody epitope and can be either linear or discontinuous. Linear epitopes consist of amino acids that are adjacent in the protein chain whereas discontinuous epitopes consist of amino acids that have been brought into proximity with one another as a result of the folding of the protein. In order to produce vaccines that contain discontinuous epitopes from multiple alleles, it is necessary to

design and express whole proteins or whole protein domains as the epitope is formed by the folding of the protein into its 3D structure. This has been done for AMA-1, an antigen that presents polymorphic, discontinuous epitopes (Remarque et al., 2008). This is in contrast to antigens that present polymorphic, linear epitopes where antigens can be designed that combine multiple epitopes in a much shorter peptide (Tetteh et al., 2005a). The latter approach would allow the inclusion of a greater number of antigen sequences without prohibitive increases in the cost of vaccine production.

1.6 Challenges of extracting repeat sequence data with modern sequencing technologies

The parallel nature of modern sequencing platforms allows for fast and inexpensive sequencing of large numbers of whole genomes. The application of these technologies to *P. falciparum* has led to a wealth of sequence data derived from thousands of clinical isolates taken from across the global distribution of the parasite. The sequence reads generated by this technology are short (50-150 base pairs). Typical approaches to analysing these short reads involve mapping reads onto a reference sequence. This alignment based approach then allows for calling of variance at the points where the mapped reads differ from the reference sequence to which they are aligned. However, this approach cannot handle highly divergent allele sequences, common to *P. falciparum* blood-stage antigens, as if a sequence is very different from the reference over the length of the sequence read it will not be mapped to the reference (MacLean et al., 2009).

This problem can be overcome by using *de novo* assembly, in which short sequence reads are joined together based on shared sequence to reconstruct the original sequence (Zerbino and Birney, 2008). This approach runs into difficulties when attempting to assemble repetitive sequence because the longer repeat and the fewer the number of changes in the repeated unit, the harder it is to determine the length of the repeat sequence (Leggett et al., 2013, Zerbino, 2010). This poses a difficulty for the use of short read sequence data in the study of the polymorphic repeat sequences that could be used to form the basis of a multi-allelic, multi-antigen vaccine.

Linear antibody epitopes are between four and 12 amino acids in length (Buus et al., 2012). It is therefore possible to elicit antibodies against an intrinsically disordered sequence solely by the inclusion of short amino acid sequences in a synthetic antigen. This thesis describes a bioinformatic method that can extract the short amino acid sequences and their frequency in a parasite population from short read sequence data for use in the computational design of synthetic antigens containing multiple potential epitopes from disordered polymorphic protein domains.

1.7 *P. falciparum* presents a wealth of potential vaccine antigens requiring characterisation and prioritisation

The *P. falciparum* genome encodes over 5500 genes (Logan-Klumpler et al., 2012) with over 90% of these producing messenger ribonucleic acid (mRNA) transcripts during the blood-stage cycle of asexual replication (Otto et al., 2010). Proteomic analysis detects almost two thousand different proteins as being present in at least one asexual parasite stage (Le Roch et al., 2004). Analysis of the proteome of lipid rafts present in just the schizont stage of the asexual life-cycle revealed the presence of over 120 proteins (Sanders et al., 2005), demonstrating the abundance of potential vaccine candidates. Selecting the relevant antigens for inclusion in a blood-stage vaccine will require analysis of the antibodies against these candidates.

Human antibodies recognise many linear epitopes from hundreds of *P. falciparum* proteins during blood-stage infection (Crompton et al., 2010a, Buus et al., 2012). Determining the contribution of these antibodies to clinical immunity is essential to designing efficacious blood-stage vaccines. Identifying key antigens for inclusion in a vaccine is a priority but it is also vital to determine how antibodies recognising these antigens work to protect the host from disease. Whilst some antibodies function by directly blocking invasion of RBCs by merozoites or cytoadherence of infected RBCs, others will induce secondary immune effector mechanisms against the parasite. Understanding the effector mechanisms of human antibodies is key to development of a malaria blood-stage vaccine (Crabb et al., 2012).

Early work used panels of mouse monoclonal antibodies raised against protein preparations or whole parasites to identify blood stage antigens (Miller et al., 1986). This approach tended to identify abundant antigens such as MSP-1 (Holder and Freeman, 1984) and MSP-2 (Miettinen-Baumann et al., 1988), although antigens of lower abundance, such as the rhoptry-associated protein RAP-1 (Clark et al., 1987), were also identified. This was followed by screening of complementary deoxyribonucleic acid (cDNA) expression libraries with serum collected from clinically immune individuals (Kemp et al., 1983), which allowed the discovery of less abundant antigens, such as the soluble antigen (S-Antigen) (Coppel et al., 1983), ring-infected erythrocyte surface antigen (RESA) (Coppel et al., 1984) and apical merozoite antigen-1 (AMA-1) (Peterson et al., 1989). This work, which focused largely on merozoite antigens, identified the majority of leading blood stage vaccine candidates, many of which have been formulated with adjuvant and trialled as vaccines (table 1.1). Advances in deoxyribonucleic acid (DNA) sequencing technology have allowed for the identification of potential antigens by detection of genetic signatures of balancing selection by the immune system (Ochola et al., 2010). The advent of whole genome sequencing means that such scans can now be carried out across the whole genome (Amambua-Ngwa et al., 2012b). Antigens such as merozoite surface protein Duffy binding like proteins (MSPBDLs) are promising candidates arising from this work, but characterisation of these new antigens is still in its infancy. Advances in the scale of sero-epidemiology and detection of genetic signatures of selection will continue to expand the list of vaccine candidates (Richards et al., 2013, Osier et al., 2014b). In order to produce optimal multi-component vaccines it is necessary to qualify these candidates by determining the efficacy of antibodies recognising them.

1.8 Correlates of protection from *P. falciparum* malaria are lacking

All blood-stage antigens that have been tested in phase II vaccine trials have been tested in animal models. Whilst studies in animal models found protection from disease, these results are not always replicated in vaccine trials or human challenge models (table 1.1). Therefore the *in vitro* and *in vivo*

use of human antibodies to qualify blood stage antigen candidates should be the focus of malaria vaccine research. Sero-epidemiology is a useful tool for the identification of vaccine candidates, but the results are often inconsistent between different cohort studies, and controlling for exposure is a persistent challenge for analysis and interpretation (Fowkes et al., 2010). The high cost of running phase II vaccine trials and human challenge models means that lab-based assays using human antibodies are needed in order to determine the potential of a vaccine candidate prior to progression into expensive validation studies. However, to date no lab based correlates of either sterile or clinical immunity to malaria have been developed.

Human antibodies recognising merozoite antigens have been observed to inhibit parasite growth in two ways. Direct inhibition can occur when antibodies prevent the invasion of merozoites through blocking binding to RBCs or to other parasite proteins or through preventing processing of antigens or by agglutination of merozoites. Antibody mediated additional mechanisms of inhibition can occur when antibodies enhance killing of parasites via opsonic phagocytosis, neutrophil respiratory burst or complement mediated lysis (Bouharoun-Tayoun et al., 1990, Bouharoun-Tayoun et al., 1995) Boyle et al., 2015).

Direct inhibition of parasite growth is the most straightforward antibody function to measure in the lab and can be standardised. Early studies using sera or purified immunoglobulin from individuals with clinical immunity to malaria found strong inhibition of *in vitro* parasite growth (Brown et al., 1983). Since this finding, growth inhibition assays (GIAs) have been used to test the capacity of serum collected from clinically immune individuals and purified antibody to directly inhibit the replication of *P. falciparum* grown in culture. The results of a longitudinal study in Mali found that increased inhibition of *in vitro* parasite growth correlated with increased protection from malaria, but that such antibodies were not sufficient to explain protection from disease (Crompton et al., 2010b). Correlation between sera inhibition of *in vitro* parasite growth and time to next infection was found in studies in Kenya, which also found that, whilst the inhibitory capacity of sera increased with age in young children, older children and adult sera had lower levels of inhibition in GIAs (Dent

et al., 2008, McCallum et al., 2008). However, the associations between growth inhibitory capacity and increased protection from malaria were not replicated in a more recent study, also conducted in Kenya (Osier et al., 2014a).

Sera from clinically immune individuals does not always inhibit parasite growth *in vitro* (Reese et al., 1981, Brown et al., 1981, Wilson and Phillips, 1976, Phillips et al., 1972, Osier et al., 2014a), implying that antibodies against *P. falciparum* can function in ways other than direct inhibition of parasite growth. Early studies demonstrated the potential for antibodies to opsonise merozoites and thus enhance phagocytosis by immune effector cells and protect from disease (Druilhe and Khusmith, 1987, Celada et al., 1982). More recently, the development of a standardised assays for phagocytosis of merozoites have been developed (Ataide et al., 2010, Hill et al., 2013) and used to demonstrate strong correlations between the ability of human sera to enhance merozoite phagocytosis and protection from disease in cohorts from Kenya, Ghana and Papua New Guinea (Osier et al., 2014a, Kana et al., 2017, Hill et al., 2013). Antibodies from clinically immune individuals have also been shown to inhibit parasite growth in the presence of monocytes (Bouharoun-Tayoun et al., 1990, Bouharoun-Tayoun et al., 1995). Whilst protocols to assay antibody dependent cellular inhibition (ADCI) are well established (Khusmith and Druilhe, 1983, Bouharoun-Tayoun et al., 1990), alterations in the assay set up influence the mechanism by which the monocytes inhibit parasite growth (Kapelski et al., 2014). In addition to phagocytosis, immune effector cells have also been shown to inhibit parasite growth by respiratory burst (Bouharoun-Tayoun et al., 1995). The capacity of sera to enhance respiratory burst of polymorphonuclear neutrophils in response to presence of merozoites has been shown to correlate with protection from malaria (Joos et al., 2010).

Whilst it is clear that cellular effector mechanisms play a key role in controlling parasite replication in the blood, one could hypothesise that, as RBCs outnumber lymphocytes 600 to one, many merozoites will escape this response by invading an RBC prior to encountering an immune effector cell. Although local inflammation and host cell suitability could act to shift the probability of encountering an immune effector cell prior to invasion, complement proteins, as soluble factors

present in the sera, will be able to act against the merozoite from the moment it is released from the schizont until it invades a new RBC and could, in theory, have a greater impact on merozoite survival and therefore parasite replication. Indeed, targeting of complement to merozoites by sera from clinically immune individuals has been shown to inhibit invasion (Boyle et al., 2015). However, difficulties in standardising this assay have prevented studies exploring the correlation with the capacity of sera to fix complement and protection from malaria. In accordance with the importance of cell and complement inhibition of merozoites being driven by antibody, cytophilic IgG3 (but not total IgG) binding to whole merozoites was shown to correlate with clinical protection from malaria (Kana et al., 2017).

Blood-stage vaccine research has focused mostly on merozoite antigens. Whilst this extracellular form of the parasite is present for only a couple of minutes of, it is the only point during the 48 hour of the asexual life cycle when the parasite is directly exposed to antibodies. However, during intraerythrocytic development the parasite exports proteins to the erythrocyte membrane, probably to facilitate cytoadherence to endothelia of microvasculature, a phenomenon that not only allows evasion of circulation through spleen, but also correlates to the aetiology of many of the symptoms of severe malaria (Miller et al., 2013). Indeed, early studies found a correlation between the capacity of serum to inhibit rosetting of RBCs, an assay for cytoadherence, and protection from future disease (Marsh et al., 1989). Antibodies recognising parasite antigens on the infected RBC surface have the potential to directly inhibit cytoadherence (Bengtsson et al., 2013), which would not only reduce severe disease symptoms but could also lead to increased clearance of infected RBCs in the spleen, or to induce cell-mediated killing of infected RBCs by immune effector cells (Lambert et al., 2014).

1.9 Human monoclonal reagents have the potential to inform design of vaccines aiming to block merozoite invasion of red blood cells

Antibodies against proteins in the rhoptries and micronemes often function by direct inhibition of parasite invasion, either through blocking the binding to RBC ligands (Persson et al., 2008) or other

parasite proteins (Maskus et al., 2016) resulting in inhibition of merozoite invasion. AMA-1 is an integral membrane protein, which is sequestered in the micronemes of the merozoite and released on contact with the RBC surface, following proteolytic processing (Narum and Thomas, 1994). Domains I and II of AMA-1 form a hydrophobic pocket (Pizarro et al., 2005, Bai et al., 2005), which binds to Rhoptry Neck Protein 2 (RON-2) (Lamarque et al., 2011, Srinivasan et al., 2011), a member of a complex of rhoptry proteins, orthologous with a family of *Toxoplasma gondii* proteins which localise to the target cell membrane following release from the rhoptries (Besteiro et al., 2009). IgG against AMA-1 has been correlated with protection from malaria (Stanisic et al., 2009, Polley et al., 2004, Gray et al., 2007, Nebie et al., 2008, Dodoo et al., 2008, Richards et al., 2013) (although not in all studies (Osier et al., 2014b)) and shown to inhibit merozoite invasion *in vitro* (Hodder et al., 2001).

Polyclonal rabbit antibodies against AMA-1 were shown to inhibit proteolytic processing of this antigen, rendering merozoites unable to invade (Dutta et al., 2003, Dutta et al., 2005), however this function has not been demonstrated with human antibodies. The study of a human monoclonal antibody demonstrated blocking of AMA-1 binding to RON-2 as the mechanism of invasion inhibition (Maskus et al., 2016), as had been suggested by work with mouse and rat monoclonal antibodies (Coley et al., 2007, Collins et al., 2007, Collins et al., 2009). Blocking of the interaction with RON-2 by human antibodies had been hypothesised due to the accumulation of polymorphic sites in the loops that surrounding the RON-2 binding pocket (Coley et al., 2006, Hodder et al., 1996). The polymorphism of AMA-1 at protective epitopes may explain the failure of single allele vaccines based on AMA-1 failing to demonstrate protection in phase II trials and human challenge models (Thompson et al., 2008, Thera et al., 2011, Laurens et al., 2017, Sheehy et al., 2012) table 1.1). Fine mapping of a rat monoclonal that inhibits merozoite invasion by binding to this region suggests that antibody recognition of the polymorphic loops is dependent on the conformation of the loop rather than linear epitopes (Collins et al., 2007). In order to induce strain transcending immune responses against AMA-1 it will therefore be necessary to combine complete AMA-1 domains of different

alleles in a single vaccine. One such vaccine, containing two alleles of AMA-1 was tested and showed limited efficacy (Sagara et al., 2009). However, vaccines containing a greater number of alleles are currently in development (Faber et al., 2016). A separate approach has also been suggested in which mutant forms of AMA-1 are engineered to direct antibody responses to conserved protective epitopes. This has been shown to induce strain-transcending inhibitory rabbit antibodies, but with a reduction in overall inhibitory capacity (Harris et al., 2014).

In addition to AMA-1, merozoites express invasion ligands from two protein families; the erythrocyte binding antigens (EBAs) sequestered in the micronemes, and the reticulocyte-binding protein homolog (RH) family proteins which are sequestered in the rhoptries. These proteins mediate invasion via binding to a range of ligands on the host cell surface (Tham et al., 2012, Bei and Duraisingh, 2012). With the exception of PfRH-5 (see below) all of these are integral membrane proteins that are variably expressed and mediate redundant invasion pathways (Lopaticki et al., 2011, Bowyer et al., 2015).

EBA-175 binds to glycophorin A on the RBC surface (Sim et al., 1994). Antibodies against EBA-175 have been shown to correlate with protection from malaria and high-density parasitaemia (Richards et al., 2010, McCarra et al., 2011, Richards et al., 2013) although this was not replicated in all studies (John et al., 2004, Okenu et al., 2000, Osier et al., 2008, Osier et al., 2014b). GIAs have demonstrated that antibodies specific for this antigen can inhibit *in vitro* parasite growth (Persson et al., 2013, Badiane et al., 2013). Naturally acquired antibodies that inhibit interaction with glycophorin-A thorough binding region II of EBA-175 were shown to correlate with protection, indicating that these antibodies act by direct blockade of merozoite-RBC interactions (Irani et al., 2015). Phase I vaccine trials with two different vaccines containing EBA-175 elicited antibodies that showed modest inhibition of parasite growth in GIA (Koram et al., 2016, Chitnis et al., 2015, El Sahly et al., 2010) (table 1.1). However, it is unlikely that a vaccination with EBA-175 alone would result in clinical immunity, as parasites are able to use other receptor-ligand interactions in place of EBA-175-

glycophorin A (Lopaticki et al., 2011, Persson et al., 2013), reflected in variable expression of this antigen and the ability of parasites to invade RBCs from which glycophorin A has been enzymatically removed (Bowyer et al., 2015).

The EBA family contains three other micronemal proteins: EBA-181, EBA-140 and EBL-1, each with its own Duffy-like binding domain (Adams et al., 1992, Adams et al., 2001), which bind glycophorins or sialic acid residues on the RBC surface (Lobo et al., 2003, Maier et al., 2003, Mayer et al., 2009, Gilberger et al., 2003, Lanzillotti and Coetzer, 2006). The reticulocyte-binding homologues (RH) are another family of invasion ligands that are sequestered in the rhoptries (Rayner et al., 2000); PfRH2, PfRH3, PfRH4 and PfRH5 bind to complement receptor 1, basigin and an unknown, trypsin resistant receptor on the RBC surface (Awandare et al., 2011, Duraisingh et al., 2003, Tham et al., 2011, Crosnier et al., 2011). Studies have found strong correlations between levels of antibodies recognising EBA and RH proteins and protection from malaria (Reiling et al., 2010, Richards et al., 2010, Richards et al., 2013). However, these findings were not for all antigens in all studies (Osier et al., 2014b, Richards et al., 2010).

Evidence from the study of human genetic variation suggests that disrupting interactions between merozoite invasion ligands and receptors on the RBC surface can protect from malaria (Leffler et al., 2017). In order to combat the redundancy of invasion pathways and variable gene expression, multiple invasion ligands will need to be incorporated into a single vaccine (Lopaticki et al., 2011). The generation of human monoclonal antibodies recognising merozoite invasion ligands will enable the detection of the protective epitopes presented by these ligands, as has been done with mouse monoclonal antibodies (Ambroggio et al., 2013), and thereby help to identify short peptide sequences for inclusion in such a vaccine.

PfRH5 is a rhoptry protein that forms a complex with *P. falciparum* PfRH5 interacting protein (PfRipr) and cysteine-rich protective antigen (CyRPA) and binds Basigin on the RBC surface and, unlike other members of the RH family, appears to be essential for parasite invasion (Baum et al., 2009, Crosnier

et al., 2011, Chen et al., 2011, Dreyer et al., 2012, Reddy et al., 2015). Human antibodies against PfRH5 have been shown to inhibit parasite invasion in GIA (Patel et al., 2013, Tran et al., 2014) and correlate with protection from disease in four out of five cohorts tested (Tran et al., 2014, Richards et al., 2013, Weaver et al., 2016, Osier et al., 2014b). Although seroprevalance was found to be low in some populations (Douglas et al., 2011, Villasis et al., 2012, Tran et al., 2014), this is not true for all populations (Richards et al., 2013, Weaver et al., 2016, Osier et al., 2014b). Rabbit and mouse antibodies induced by a 3D7 PfRH5 delivered on a viral vector were able to inhibit growth of both homologous and a heterologous strain (Douglas et al., 2011, Douglas et al., 2014), due to the limited polymorphism present in the gene encoding PfRH5, which may be due to restrictions imposed by binding to Basigin (Hayton et al., 2008, Wanaguru et al., 2013). Immunization with a PfRH5 viral vectored vaccine protected *Aotus* monkeys from challenge with heterologous parasites and produced antibodies that inhibit parasite growth *in vitro* (Douglas et al., 2015). PfRH5 has been formulated as a viral-vectored vaccine with chimpanzee adenovirus 63 (ChAd63) and modified vaccinia virus Ankara (MVA) strain and tested in a human challenge model (trial number NCT02181088), although at the time of writing the results of this trial have not been published. GIAs and epitope mapping using mouse monoclonal antibodies have determined two linear epitopes recognised by inhibitory antibodies (Douglas et al., 2014). Confirmation that these epitopes also elicit functional human antibodies would mean that these two short peptides could be included in a multi-antigen vaccine.

Multiple merozoite surface proteins (MSPs) contain epidermal growth factor-like (EGF) domains that may mediate initial contact with the target RBC membrane (Goel et al., 2003, Kariuki et al., 2005, Boyle et al., 2010, Puentes et al., 2005, Puentes et al., 2003); antibodies against these domains can inhibit invasion, presumably by directly blocking this interaction (Maskus et al., 2015). MSP-1, the most abundant protein on the merozoite surface (Gilson et al., 2006), is cleaved into four fragments that remain associated on the surface of the merozoite (Holder et al., 1987, McBride and Heidrich, 1987), although only the C-terminal 42 kilodalton fragment (MSP-1₄₂) is covalently linked to the GPI

anchor. During invasion MSP-1₄₂ is cleaved, releasing the N-terminal fragments and leaving just a 19 kilodalton fragment (MSP-1₁₉) that is carried into the RBC (Blackman and Holder, 1992, Stafford et al., 1994, Blackman et al., 1990).

Antibodies against MSP-1₁₉ have been found to correlate with protection in some cohorts (Egan et al., 1996, Perraut et al., 2005, Stanisic et al., 2009), although a greater number of studies in different cohorts found no correlation (Egan et al., 1996, Cavanagh et al., 2004, Conway et al., 2000, Doodoo et al., 2008, Nebie et al., 2008, Osier et al., 2008, Soe et al., 2004, Richards et al., 2013). Only two studies, both conducted in Papua New Guinea have analysed antibodies against MSP-1₄₂ and protection from malaria with one study reporting an association (al-Yaman et al., 1996) and another reporting no association (Richards et al., 2013). Mouse and rabbit antibodies against MSP-1₁₉ and MSP-1₄₂ have been shown to directly inhibit parasite growth *in vitro* (Blackman et al., 1990, Chappel and Holder, 1993, Bergmann-Leitner et al., 2006). Whilst this has not been directly demonstrated for human antibodies, mutation of MSP-1₁₉ EGF domains does rescue inhibition by human sera (O'Donnell et al., 2001, Murhandarwati et al., 2008). Antibodies against MSP-1₄₂ could be blocking the proposed binding of heparin-like polysaccharides (Boyle et al., 2010) or Band-3 (Goel et al., 2003, Kariuki et al., 2005) on the RBC surface by MSP-1₄₂. Mouse antibodies and human sera from exposed children have been shown to prevent MSP-1₄₂ processing (Guevara Patino et al., 1997, Nwuba et al., 2002), however, this is not considered to be major functional mechanism of naturally acquired antibodies (Moss et al., 2012). Furthermore, IgG against MSP-1₄₂ does not correlate with the invasion inhibitory capacity of sera from children living in an endemic area of Papua New Guinea (McCallum et al., 2008).

Vaccination with MSP-1₁₉ was shown to be protective in a non-human primate model, although only when used with Freund's adjuvant which is not licensed for use in humans (Egan et al., 2000, Kumar et al., 1995, Kumar et al., 2000). In the same model, vaccination with MSP-1₄₂, formulated with Montanide ISA-720, an adjuvant approved for human use, was shown to be protective against

challenge with homologous parasites (Lyon et al., 2008). MSP-1₁₉ has been included in several vaccine formulations although it often proves to be poorly immunogenic (Keitel et al., 1999, Chitnis et al., 2015) and has not progressed to phase II trials (table 1.1). In a formulation in which MSP-1₁₉ elicited antibodies in a majority of participants, these did not inhibit parasite growth *in vitro* (Hu et al., 2008, Malkin et al., 2008). Two different alleles of MSP-1₄₂ have been formulated as two separate recombinant vaccines; both were immunogenic but only vaccination with the 3D7 allelic form produced antibodies capable of inhibiting parasite growth *in vitro* (Otsyula et al., 2013, Ockenhouse et al., 2006). The phase II vaccine trial with this allelic form did not demonstrate any protection from disease (Ogutu et al., 2009). This antigen has also been included in a viral vectored vaccine in combination with conserved N-terminal MSP-1 peptide sequences. Whilst this vaccine did induce antibodies recognising MSP-1₄₂, these did not inhibit parasite growth *in vitro* or in a human challenge infection model (Sheehy et al., 2011, Sheehy et al., 2012).

MSP-10 was more recently identified based on homology with MSP-1 (Black et al., 2003) and has been implicated in binding the RBC membrane (Puentes et al., 2005). Three human monoclonal antibodies isolated from a clinically immune individual recognising conformational epitopes in the two EGF domains of MSP-10 have been shown to directly inhibit parasite invasion (Maskus et al., 2015).

Ring-infected surface antigen (RESA) is localised to the micronemes of the merozoite but appears to only be released following invasion (Brown et al., 1985) whereupon it localises to the membrane of the host cell and interacts with spectrin (Foley et al., 1991). High levels of anti-RESA antibodies correlate with lower parasite densities (Petersen et al., 1990) and protection from malaria (Astagneau et al., 1994a, Astagneau et al., 1994b, Astagneau et al., 1995). However, levels of IgG1 to a repeat unit of RESA were found to be negatively correlated with protection in one study (Dubois et al., 1993). High levels of IgG2 were shown to correlate with protection from disease in a population with a high prevalence of the H131 allele, suggesting a role for antibodies against RESA in immune

clearance (Aucan et al., 2000). The inhibitory capacity of sera collected from clinically immune individuals was found to correlate with levels of RESA specific IgG (Wahlin et al., 1984, Berzins et al., 1986) and human and rabbit antibodies recognising an eight amino acid repeat unit of RESA inhibited parasite growth *in vitro* (Berzins et al., 1986). The first described human monoclonal antibodies recognising a *P. falciparum* antigen were against RESA and were also shown to inhibit parasite growth *in vitro* (Udomsangpetch et al., 1986, Berzins et al., 1985, Berzins et al., 1986). These results are confusing, given the description of apparent internal localisation of the protein in the merozoite and the fact that it was elsewhere shown that RESA does not appear to be essential for parasite growth (Cappai et al., 1989). However, the fact that antibodies against the repeat sequences of RESA elicited by immunisation of *Aotus* monkeys were found to protect these animals from infection with *P. falciparum* (Collins et al., 1986) encouraged the inclusion of the constant domain of this antigen in combination B, the first multi-antigen *P. falciparum* vaccine to be trialled. Results of a phase II trial of combination B vaccine showed reduced parasite burden in vaccinated individuals (Genton et al., 2002).

1.10 Antibodies against antigens on the merozoite surface can trigger secondary immune mechanisms leading to merozoite neutralisation

Antibodies recognising proteins present on the surface of the merozoite surface typically do not directly inhibit merozoite invasion (Beeson et al., 2016). However, antibodies against many of these proteins have been found to correlate with protection from disease (Richards et al., 2013, Osier et al., 2014b). Furthermore, assays of immune effector mechanisms have shown that antibodies against MSPs can enhance destruction of merozoite (Boyle et al., 2015, Osier et al., 2014a).

Population genetic analysis indicates that a highly-polymorphic N-terminal region of MSP-1, termed MSP-1 block 2, is under strong immune pressure (Conway et al., 2000, Polley et al., 2003a).

Accordingly, antibodies, especially IgG3, recognising this region have been found to associate with protection from malaria in several populations (Conway et al., 2000, Cavanagh et al., 2004, Polley et

al., 2003b). This protection is particularly associated with the polymorphic repeat sequences found in some alleles (Polley et al., 2003b). However, these results have not been replicated in every cohort (Osier et al., 2008, Gray et al., 2007). Human antibodies against MSP-1 block 2 do not inhibit parasite growth directly but have been shown to inhibit merozoites in the presence of monocytes in a strain specific manner (Galamo et al., 2009). Rabbit antibodies against MSP-1 block 2 were shown to inhibit merozoite invasion in the presence of complement in an allele specific manner (Boyle et al., 2015). Vaccination with a recombinant protein based on MSP-1 block 2 protected two out of four Aotus monkeys in a non-human infection challenge model (Cavanagh et al., 2014).

MSP-2 is the second most abundant protein on the merozoite surface (Sanders et al., 2005) and is dimorphic (Fenton et al., 1991). Naturally acquired IgG3 against MSP-2 were found to correlate with protection (Taylor et al., 1998, Metzger et al., 2003, Polley et al., 2006, Stanisic et al., 2009, Osier et al., 2008, Flueck et al., 2009, Osier et al., 2014b). This was not shown in all cohorts (Sarr et al., 2006, Polley et al., 2006, Richards et al., 2013) and was found to be allele specific in at least one (Scopel et al., 2007). Vaccine induced antibodies against MSP-2 have been shown to inhibit parasite growth in ADCl assays (McCarthy et al., 2011). Naturally acquired MSP-2 antibodies were shown to enhance phagocytosis of merozoites (Osier et al., 2014a) and rabbit antibodies recognising one allele of this antigen have been shown to strongly inhibit invasion of merozoites in the presence of human complement (Boyle et al., 2015). Mutation of the Fc region that reduces binding to Fc receptors has been used to demonstrate inhibition of parasite growth via opsonic phagocytosis, ADCl and activation of complement by a recombinant human monoclonal recognising MSP-2 (Stubbs et al., 2011, Boyle et al., 2015).

MSP-3 was the first member identified of a family of six peripheral membrane proteins, the MSP3/6 family, that have a common N-terminal motif and conserved C-terminal domains and are expressed on the surface of the merozoite (Oeuvray et al., 1994, Singh et al., 2009). The function of these proteins remains unknown, although binding of the RBC membrane has been demonstrated for one

family member (Sakamoto et al., 2012). Truncation of MSP-3 resulted in a reduction of invasion efficiency, suggesting that this family has a role in merozoite invasion but that there are overlapping functionalities (Mills et al., 2002). Antibodies against MSP-3 have been found to correlate with protection in many studies (Richards et al., 2013, Meraldi et al., 2004, Nebie et al., 2008, Osier et al., 2014b, Osier et al., 2008, Osier et al., 2007, Polley et al., 2007) and this has also been shown for several other MSP-3/6 family members (Richards et al., 2013). In some studies the association between anti-MSP-3 antibodies and protection was found to be allele specific (Osier et al., 2007, Osier et al., 2008) and one study found no correlation (Gray et al., 2007). Human antibodies against MSP-3 also inhibit parasite growth in the presence of monocytes (Oeuvray et al., 1994) and can enhance phagocytosis of merozoites (Osier et al., 2014a). The analysis of a human monoclonal antibody binding to one heptad repeat unit of MSP-3, produced both as IgG1 and IgG3, showed inhibition of parasite growth via ADCC for both antibody isotypes (Lundquist et al., 2006). The conserved C-terminal region of MSP-3 was formulated as a vaccine that induced antibodies capable of inhibiting parasite growth *in vitro* in the presence of monocytes and *in vivo* in a mouse model (Druilhe et al., 2005). Subsequently, a small scale trial of this vaccine demonstrated protection (Sirima et al., 2011). Antibodies against each of the MSP-3/6 family members inhibited parasite growth in the presence of monocytes and were shown to be cross-reactive (Singh et al., 2009). Production of human monoclonal reagents against MSP-3/6 family members will aid the design of vaccines based on identification of common functional epitopes.

Glutamate rich protein (GLURP) is another peripheral membrane protein found on the surface of merozoites, which contains two polymorphic repeats, R₁ and R₂ and a conserved non-repeat region, R₀ (Borre et al., 1991, Hogg et al., 1993). High levels of antibodies against GLURP correlate with low density parasitaemia (Hogg et al., 1992) and protection from disease (Nebie et al., 2008, Dodoo et al., 2008, Meraldi et al., 2004), although this was not found for antibodies recognising GLURP R₂ in one cohort (Richards et al., 2013). Human antibodies against GLURP have been shown to inhibit parasites in the presence of monocytes (Theisen et al., 1998) and such antibodies were induced in

adults by vaccination with a section of the R₀ peptide. This region of GLURP combined with the conserved region of MSP-3 in a hybrid vaccine that showed modest and not statistically significant efficacy in a phase II trial (Sirima et al., 2016).

Serine repeat antigen-5 (SERA-5) is a member of a family of SERA proteins (Bourgon et al., 2004). SERA-5 is a soluble, exported protein expressed during the late trophozoite and schizont stage of the asexual cycle (Delplace et al., 1987, Knapp et al., 1989). The N-terminal domain of SERA-5 consists of a polymorphic repeat sequence whereas the C-terminal domain is conserved (Morimatsu et al., 1997, Fox et al., 1997). SERA-5 undergoes proteolytic processing during schizont maturation and an N-terminal fragment remains associated with merozoites following schizont rupture and egress from erythrocytes (Li et al., 2002). Antibodies recognising this N-terminal fragment correlated with lower parasitaemia amongst adults in an endemic area of Brazil (Banic et al., 1998) and signs of clinical immunity in adults and children living in Uganda (Okech et al., 2001). Accordingly, antibodies recognising SERA-5 have been shown to correlate with protection from malaria (Richards et al., 2013, Okech et al., 2006). In combination with monocytes, antibodies against SERA-5 can inhibit parasite growth (Soe et al., 2002). The phase Ib trial of a recombinant vaccine containing the N-terminus of a single allele of SERA-5 has shown protection from clinical malaria amongst high responders (Yagi et al., 2016). The predicted disorder of both GLURP and SERA proteins means that identification of protective epitopes by analysis of human monoclonal antibodies would allow the use of short peptide sequences within immunogenic constructs to induce functional antibody responses.

1.11 Antibodies target destruction of infected red blood cells

Similar immune effector mechanisms appear to be involved in inhibition of parasite growth following invasion of the RBCs by merozoites; complement and *P. falciparum* specific antibodies have been shown to enhance killing of infected RBCs via respiratory burst by neutrophils (Salmon et al., 1986) and phagocytosis (Celada et al., 1983, Kumaratilake et al., 1997). Whilst the intra-erythrocytic

parasite is shielded inside the RBC membrane, this stage of the parasite does express variant antigens, namely *P. falciparum* erythrocyte membrane protein-1 (PfEMP-1) and rifins, on the surface of the infected RBC in order to adhere to other RBCs or endothelial cells thus preventing clearance in the spleen (Carlson et al., 1990, Carlson et al., 1994, Chen et al., 1998, Miller et al., 2013, Rowe et al., 1995, Rowe et al., 1997, Scherf et al., 2008, Vigan-Womas et al., 2012, Goel et al., 2015). These proteins are encoded by gene families comprising ~60 or ~150 members for PfEMP-1 and rifins respectively (Fernandez et al., 1999, Horrocks et al., 2004). This has previously discouraged development of vaccines based on these antigens, apart from in the special case of pregnancy associated malaria. The fact that antibodies against the PfEMP-1 variant (VAR2-CSA) that causes pregnancy associated malaria protect against disease (Ampomah et al., 2014) has encouraged the development of a viral vectored vaccine based on this variant (Andersson et al., 2017), the phase I clinical trial for which is currently underway (trial number NCT02647489). Interestingly, increased phagocytosis of infected RBCs appears to be a factor contributing to resistance to malaria conferred by human genetic polymorphisms affecting the production of alpha (alpha-thal 1 trait ($\alpha\alpha$), alpha-thal 2 trait ($-\alpha/-\alpha$), Hb H/Hb Constant Spring (CS) ($-\alpha/\alpha_{CS}$), HB CS trait ($\alpha\alpha/\alpha\alpha_{CS}$) and CS disease ($\alpha\alpha_{CS}/\alpha\alpha_{CS}$) or beta (beta+ thalassemia trait (HbA_{39 (C>T)}/HbA)) haemoglobin chains or introducing a mutation in the beta haemoglobin chain (sickle cell trait (HbS/HbA) which are common in malaria endemic regions (Yuthavong et al., 1990, Ayi et al., 2004). Additionally infected RBCs with erythrocyte glucose-6-phosphate dehydrogenase (EG6PD) deficiency, caused by deletion of the EG6PD gene on the X chromosome, were shown to be phagocytosed at the ring stage rather than trophozoite stage of intraerythrocytic development, which reduces the toxicity to the phagocytosing cells (Cappadoro et al., 1998).

Antibodies recognising a range of type A PfEMP-1 antigens were raised in rabbits and shown to inhibit rosetting and enhance phagocytosis of infected RBCs (Ghumra et al., 2012). Broad-specificity anti-PfEMP-1 antibodies were also observed following controlled infection in naïve adults, leading to the hypothesis that multiple *var* genes are expressed in the early stage of infection in order to allow

the parasite to find a permissive niche in the host circulatory system (Turner et al., 2011). This suggests that a vaccine comprising multiple PfEMP-1 motifs could induce broad immune responses against this antigen (Bull and Abdi, 2016). The recent discovery in two separate Kenyan populations of monoclonal antibodies, containing a large, mutated insertion from another gene, with broad specificity for rifins does lend promise to the development of a vaccine against these highly variant gene families (Tan et al., 2016). However, engineering a vaccine that can induce rare antibodies is not straightforward; a rare, broadly neutralizing antibody against human immunodeficiency virus (HIV) glycoprotein 120 was discovered over twenty years ago (Trkola et al., 1996), however, as yet no vaccine strategies have been developed to elicit antibodies with similar specificity and potency (Scanlan et al., 2007).

1.12 Recombinant human monoclonal antibodies enable the assaying of antibody mediated killing of *P. falciparum* merozoites and infected red blood cells induced by antibody responses to specific antigens

The production of recombinant monoclonal antibody reagents allows for the introduction of mutations that can remove C1q and Fc binding sites. Such reagents provide the perfect controls for assaying antibody mediated complement and cell driven parasite inhibition (Stubbs et al., 2011, Boyle et al., 2015). Improved standardisation of these assays will support key findings that will contribute to vaccine design. It is possible that antibody mediated enhancement of complement and cellular immune effector mechanisms is only optimal in the presence of multiple antibody specificities (Boyle et al., 2015). In order to explore the minimal number of antigen epitopes that need to be recognised in order to enhance these immune effector mechanisms it will be necessary to generate large numbers of human monoclonal antibody reagents.

Advances in whole genome sequencing, proteomics and recombinant protein expression, have enabled the study of signatures of immune selection, protein localisation and antibody responses for large numbers of antigens. This has led to the identification of a number of novel vaccine candidates (Osier et al., 2014b, Sanders et al., 2005, Crompton et al., 2010a, Amambua-Ngwa et al., 2012b). The

generation of human monoclonal reagents against these antigens will help to determine which elicit functional antibody responses.

1.13 MSP-1 is processed before and during merozoite invasion and encodes tripeptide repeat sequences

The most abundant antigens on the merozoite surface are the peptides of the MSP-1 complex, which is conserved amongst all *Plasmodium* species studied (Cooper, 1993) and is the main focus of this study. The MSP-1 precursor is a large (~190 kDa) protein that forms a complex with a tetramer of MSP-6 proteins and MSP-7 (Kauth et al., 2003, Kauth et al., 2006, Lin et al., 2014, Pachebat et al., 2001, Trucco et al., 2001). This complex is localised to the merozoite membrane via a C-terminal GPI anchor (Gerold et al., 1996). MSP-1 has been the focus of much research due to its capacity to antibodies able to inhibit *in vitro* parasite growth (Holder, 2009). The inability to knockout MSP-1, despite multiple attempts, suggests that this protein is essential for parasite growth (Combe et al., 2009, Drew et al., 2004, O'Donnell et al., 2000) but also means we do not have direct evidence for the function of the protein (Das et al., 2015). There is indirect evidence that MSP-1 plays a role in invasion as it has been shown to bind to glycophorin A (Baldwin et al., 2015, Su et al., 1993), Band 3 (Goel et al., 2003, Li et al., 2004) and heparin-like molecules (Boyle et al., 2010, Zhang et al., 2013), all molecules present on the RBC surface. The ability of heparin or heparin-like polysaccharides to block merozoite invasion suggests that interaction between merozoite surface proteins and heparin is an essential step in the invasion process (Boyle et al., 2010, Zhang et al., 2013, Clark et al., 1997, Crick et al., 2014, Kulane et al., 1992). However, the role played by MSP-1 in invasion is yet to be determined.

The MSP-1/6/7 complex undergoes proteolytic processing as the merozoite egresses from the schizont and invades a new RBC. MSP-7 is processed prior to incorporation into the MSP-1/6/7 complex (Kauth et al., 2006). Minutes before parasite egress, subtilisin-1 is released into the PV which cleaves MSP-1 at three sites to produce four peptides (MSP-1₈₃, MSP-1₃₀, MSP-1₃₈ and MSP-

1₄₂) that remain associated to the merozoite surface via the carboxy-terminal (C-terminal) GPI anchor of MSP-1₄₂ (Freeman and Holder, 1983, Blackman, 2000, Koussis et al., 2009; figure 1.4). Subtilisin-1 also truncates MSP-6 and MSP-7 (Koussis et al., 2009). This processing is required for efficient parasite egress, seemingly via enabling binding of the host cell cytoskeleton (Das et al., 2015). Prior to invasion, subtilisin-2 cleaves MSP-1₄₂ at a single site (Harris et al., 2005) releases the MSP-1 complex from the merozoite membrane (Blackman et al., 1991, Riglar et al., 2011) leaving the MSP-1₁₉ to be taken into the host erythrocyte (figure 1.4).

The MSP-1 sequence has been divided into 17 blocks based on sequence polymorphism (Tanabe et al., 1987). MSP-1 block 2, contained within the MSP₈₃ processing fragment, is one of the most divergent regions of the antigen (Miller et al., 1993). Three allelic types of MSP-1 block 2 have been defined: 3D7 like, MAD20 like and RO-33 like (Miller et al., 1993). Different forms of degenerate tripeptide repeats are found in the 3D7 and MAD20 allelic types but the repeat sequence is not present in the RO-33 type allele (Miller et al., 1993).

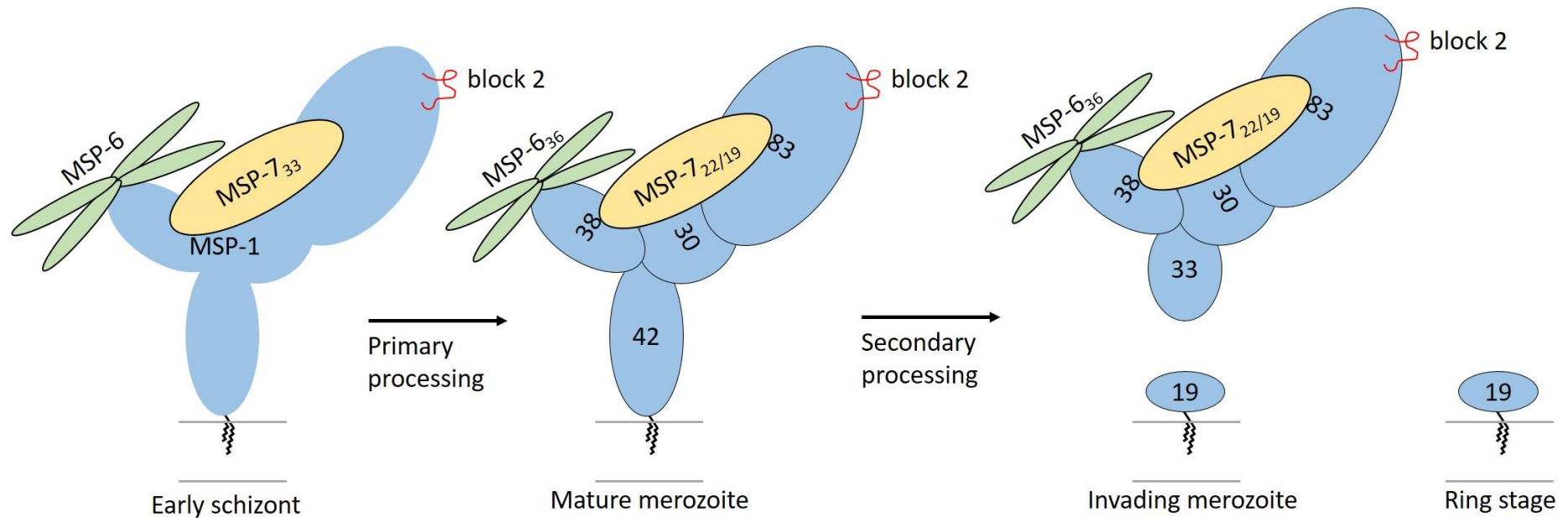


Figure 1.4 Schematic representation of processing of MSP-1. MSP-1 is expressed as a 190 kDa protein (blue) with the MSP-1 block 2 sequence (red) located near the N-terminus. This protein is associated to the merozoite membrane by a C-terminal GPI anchor. MSP-1 forms a complex with a tetramer of MSP-6 (green) and MSP-7₃₃ (yellow) a fragment resulting from processing prior to complex formation with MSP-1. Minutes before egress from the host RBC the MSP-1/6/7 complex is processed. MSP-1 is cleaved at three sites by subtilisin-1 to form four fragments (MSP-1₈₃, MSP-1₃₀, MSP-1₃₈ and MSP-1₄₂) that remain associated. MSP-6 is cleaved and the resulting MSP-6₃₆ fragment remains associated to MSP-1. Alternate cleavage of MSP-7₃₃ results in either an MSP-7₂₂ or MSP-7₁₉ fragment which also remains in the MSP-1/6/7 complex. After egress from the schizont, extracellular processing of the MSP-1/6/7 complex by subtilisin-2 cleaves MSP-1₄₂ into two fragments; MSP-1₁₉ that stays associated to the merozoite membrane and enters the RBC invaded by the merozoite and MSP-1₃₃ which detaches from the merozoite membrane along with the rest of the MSP-1/6/7 complex. Figure adapted from Koussis et al., 2009.

1.14 Aims and objectives


Development of the next generation of *P. falciparum* vaccines will be aided by use of techniques that can access short read sequence data arising from polymorphic repetitive regions of antigen genes. It will also be benefited by *in silico* approaches that allow this sequence data to be used to inform multivalent vaccine design. The aims of this work are to develop tools for bioinformatic analysis of repetitive antigen sequences from short read data and *in silico* design of multivalent antigen constructs based on these repeat sequences. This project also aims to trial a technique for monoclonal antibody production that would enable the testing of efficacy of naturally occurring antibodies recognising novel vaccine antigens.

The work focuses on MSP-1 block 2 as this region of MSP-1 encodes a highly polymorphic repeat sequence (Miller et al., 1993) which is well characterised. To meet the first aim of this work, a library of known MSP-1 block 2 sequences was constructed by mining of sequence databases. This library was used to generate short sequence reads for use in validation of two approaches for extracting sequence data; the first was *de novo* assembly of short read sequences, which required optimisation of *de novo* assembly algorithms and validation using data generated from known MSP-1 block 2 sequences. The second approach to extracting sequence data was to align short reads to a reference library of sequences representing the range of known MSP-1 block 2 sequences. This approach involved the construction and validation of the reference library. Both approaches were applied to short read sequence data from the Pf3k project. The approaches are compared to each other and to MSP-1 block 2 sequences from long read data.

To meet the second aim of this thesis, algorithms were developed to rationally design multivalent antigens based on the MSP-1 block 2 repeat sequences identified in the previous work. The output of these algorithms are analysed *in silico* for coverage of MSP-1 block 2 sequences and compared to previous work in which multivalent antigens were designed by eye (Tetteh and Conway, 2011).


To meet the third aim of this thesis MSP-1 tetramers were created to demonstrate the potential use of such tetramers for the isolation of memory B-cells from the blood of malaria exposed individuals. mRNA encoding the B-cell receptor was amplified to allow for sequencing of the Ig genes , which would enable the production of the antibody specificity encoded by the B-cell.

Table 1.1. Results of clinical trials of blood stage *P. falciparum* vaccines.

Antigen(s)	Vaccine	Phase	Participants	Results	Reference(s)
MSP-1 ₁₉	P30P2 MSP-119	I	16 naïve adults	Poor immunogenicity	(Keitel et al., 1999)
MSP-1 ₄₂ (3D7 allele)	FMP001	I	15 naïve adults	Immunogenic (cellular and antibody responses). Antibodies inhibit parasite growth in GIA	(Ockenhouse et al., 2006)
		I	40 exposed adults	Boosting of antibody responses to homologous and heterologous MSP-1	(Thera et al., 2006)
		I	40 exposed adults	Boosting of MSP-1 ₄₂ antibody responses	(Stoute et al., 2007)
		Ib	135 exposed children (1-4 years)	Boosted antibody responses	(Withers et al., 2006)
		II	400 exposed children (1-4 years)	Boosted specific antibodies; no protection from disease	(Ogutu et al., 2009)
MSP-1 ₄₂ (FVO allele)	FMP010	I	58 naïve and exposed adults	Boosting of antibody responses to homologous and heterologous MSP-1; no change to inhibitory capacity of sera in GIA	(Otsyula et al., 2013)
MSP-1 ₄₂ + blocks 1,3,5 and 12 (3D7 allele)	ChAd63-MVA MSP1 	Ia	16 naïve adults	Immunogenic (cellular and antibody responses). Antibody titres not high enough to inhibit in GIA	(Sheehy et al., 2011)


		Ila	10 naïve adults	Immunogenic but not protection in human challenge model	(Sheehy et al., 2012)
MSP-2 central domain (3D7 allele) + CSP	Ro 46-2717 + Ro 46-2924	I	39 naïve adults	Moderate immunogenicity (humoral) no significant difference in human challenge model (5 volunteers) compared to unvaccinated.	(Sturchler et al., 1995)
MSP-3 C-terminal domain (conserved)	MSP3-LSP	I	35 naïve adults	Immunogenic (antibody and cellular); elicited cytophilic antibody response capable of inhibiting parasite growth in ADCI and <i>in vivo</i> in a mouse model	(Audran et al., 2005, Druilhe et al., 2005)
		Ib	30 exposed adults	No detection of boosting of antibody (high baseline); suggested increased cellular response.	(Sirima et al., 2007)
		Ib	45 exposed children (1-2 years)	Protection from clinical malaria	(Sirima et al., 2011)
GLURP ₈₅₋₂₁₃	GLURP ₈₅₋₂₁₃ LSP	I	36 naïve adults	Induced mainly cytophilic antibody response that showed parasite growth inhibition in presence of monocytes	(Hermesen et al., 2007)
AMA-1 (3D7 allele)	AMA1	I	29 naïve adults	Poor immunogenicity	(Saul et al., 2005)

AMA-1 (3D7 allele)	AMA1-C1	I	72 naïve adults	Addition of CPG7909 adjuvant increased AMA-1 specific antibody titres. IgG inhibited homologous parasite growth in GIA	(Mullen et al., 2008)
AMA-1 (FVO allele)	PfAMA1-FVO	I	56 naïve adults	Immunogenic (cellular and antibodies). Antibodies inhibit invasion in GIA	(Roestenberg et al., 2008)
Loop 1 of AMA-1 domain III	PEV301	Ia	20 naïve adults	Immunogenic	(Genton et al., 2007)
Loop 1 of AMA-1 domain III + CSP + TRAP	FFM ME-TRAP+PEV3A	I/IIa	24 naïve adults	Immunogenic (antibodies). Reduction of parasite growth rate versus controls but no significant reduction in time to detectable parasitaemia found in human challenge model	(Thompson et al., 2008)
AMA-1 (3D7 allele)	FMP2.1	I	23 naïve adults	Immunogenic (cellular and antibodies). Antibodies inhibition parasite growth in GIA and block processing AMA-1	(Polhemus et al., 2007)
		I	60 exposed adults	Anti-AMA-1 antibodies significantly boosted by vaccination. No significant change in growth inhibitory capacity of sera	(Thera et al., 2008)


		I	33 exposed children (2-3 years)	Anti-AMA-1 antibodies significantly boosted by vaccination.	(Dicko et al., 2008)
		II	400 exposed children (1-6 years)	Limited vaccine efficacy (20%) but increased efficacy (68%) against infection with parasites bearing homologous AMA-1	(Thera et al., 2011)
		II	300 exposed children	IgG from vaccinated individuals showed increased inhibition of homologous parasite growth, but no correlation with protection from malaria	(Laurens et al., 2017)
AMA-1 (FVO allele)	ChAd63-MVA AMA-1 	Ila	9 naïve adults	Immunogenic but no protection in human challenge model	(Sheehy et al., 2012)
SERA-5 (Honduras allele)	BK-SE36	Ia	40 naïve adults	Immunogenic (antibody responses)	(Tanabe et al., 2010)
		Ib	122 exposed children and adults (6-40 years)	Immunogenic in sero-negatives; protection from infection especially in high responders	(Palacpac et al., 2013, Yagi et al., 2016)
EBA-175 Region II	EBA-175-RII-NG	I	18 naïve adults	Elicits EBA-175 specific antibodies that show modest parasite growth inhibition in GIA	(El Sahly et al., 2010)

		I	52 exposed adults	Boosts EBA-175 specific antibodies that increase growth inhibition in GIA compared to controls	(Koram et al., 2016)		
MSP-1 block 1 peptide + PF3D7_1026300 peptide + AARP-1 peptide + CSP NANP repeats	SPf66*	II	1548 adults and children resident in low endemic area	33% vaccine efficacy	(Valero et al., 1993)		
		II	537 adults and children resident in low endemic area	67 % vaccine efficacy	(Sempertegui et al., 1994)		
		II	1257 adults and children resident in low endemic area	35 % vaccine efficacy	(Valero et al., 1996)		
		II	572 adults and children resident in low endemic area	No efficacy	(Urdaneta et al., 1998)		
		II	586 children (1-5 years)	31% vaccine efficacy	(Alonso et al., 1994)		
		II	630 children (6-11 months)	No significant efficacy	(D'Alessandro et al., 1995)		
		II	150 children (6-11 months)	No efficacy	(Bojang et al., 1997)		
				I	69 adults	No impact on MSP-1 allele frequency in vaccinated individuals	(Masinde et al., 1998)
				IIb	1207 children (one month)	2% vaccine efficacy	(Acosta et al., 1999)
		IIb	1348 children (2-15 years)	No efficacy	(Ballou et al., 1995)		

AMA-1 (3D7 allele) + MSP-1 (Palo Alto strain) + SERA (FCR3 allele) + CSP + LSA-1 + TRAP +Pfs25	NYVAC-Pf7 	I/IIa	35 naïve adults	Poor immunogenicity (antibodies). Good cellular immune response. Protected one individual from infection and caused slight but significant delay in time to parasitaemia.	(Ockenhouse et al., 1998)
MSP-1 blocks 3 and 4 (K1 allele) + MSP-2 (3D7 allele) + RESA constant domain (FQ-27/PNG allele)	Combination B	I	36 naïve adults	Elicited humoral and cellular responses against all three antigens.	(Saul et al., 1999)
		I	10 exposed adults	Moderate boosting of antibody levels (high baseline); cellular responses to MSP-1 and RESA.	(Genton et al., 2000)
		IIa	12 naïve adults	Strong cellular but weak humoral response. No significant difference in immune response or initial parasite growth rates following controlled infection with homologous strain.	(Lawrence et al., 2000)
		I-IIb	120 children 5-9 years	Boosting of humoral response to all three antigens; cellular response to MSP-1 only. Reduction of parasite burden; vaccinated individuals had lower prevalence of parasites with MSP-2 vaccine allele.	(Genton et al., 2002, Genton et al., 2003)

AMA-1 domain III + MSP-1 ₁₉ (3D7 allele)	PfCP-2.9	I	72 naïve adults	Immunogenic but no inhibition of parasite growth in GIA.	(Hu et al., 2008, Malkin et al., 2008)
MSP-1 ₄₂ + blocks 1,3,5 and 12 (3D7 allele) + AMA-1 (FVO allele)	ChAd63-MVA MSP-1 +ChAd63-MVA AMA-1 	Ila	9 naïve adults	Immunogenic but no protection in human challenge model.	(Sheehy et al., 2012)
MSP3 C-terminal domain + GLURP R ₀ non-repeat region	GMZ2	Ia	30 naïve adults	Elicited antibodies and memory B-cells recognising MSP-3 and GLURP that persisted for up to one year post vaccination.	(Esen et al., 2009)
		I	40 exposed adults	Significant boosting of antibody and memory B-cells specific for MSP-3 and GLURP despite high baseline.	(Mordmuller et al., 2010)
		Ib	30 exposed children (1-5 years of age)	Boosted antibody responses to both MSP-3 and GLURP; elicited antigen-specific memory B-cells that persisted for up to one year post vaccination.	(Belard et al., 2011)
		IIb	1849 exposed children (1-5 years)	Vaccine efficacy of 14%; vaccine efficacy increased with age and levels of vaccine specific antibodies but was not statistically significant.	(Sirima et al., 2016)

EBA-175 (Camp allele) + MSP-1 ₁₉ (FVO allele)	JAIVAC-1	I	45 naïve adults	Elicits antibody responses to EBA-175 but not MSP-1 ₁₉ . IgG inhibit growth of parasites expressing homologous EBA-175.	(Chitnis et al., 2015)
MSP-2 (3D7 + FVO alleles)	MSP2-C1	I	36 naïve adults	Adverse effects in higher dose; elicited antibodies recognising both MSP-2 alleles that inhibited parasite growth in the presence of monocytes/	(McCarthy et al., 2011)
AMA-1 (3D7 and FVO alleles)	AMA1-C1	II	300 exposed children	No reduction in parasite densities amongst vaccinated individuals	(Sagara et al., 2009)
MSP-1 ₄₂ (3D7 and FVO alleles) + AMA-1 (3D7 and FVO alleles)	BSAM2	I	30 naïve adults	Some systemic adverse reactions; elicited antibodies against all antigens; total IgG inhibitory in GIA against homologous parasite strains	(Ellis et al., 2012)

The 43 clinical trials of vaccines containing one or more blood stage antigen that have been published to date are shown. The antigens and (where relevant) alleles present in the vaccine are listed along with the phase of the trial and summaries of participants and results. The table is coloured by the number of blood stage antigens/alleles: yellow denotes a single allele of a single antigen; blue denotes single alleles of multiple antigens; green denotes multiple alleles of a single antigen; and mauve denotes multiple alleles and multiple antigens. Darker shades highlight trials in which vaccine efficacy has been assessed, either by protection from disease by natural infection or human challenge model. Viral vectored vaccines are indicated (). Abbreviations: AMA-1 – apical membrane antigen-1; MSP – merozoite surface protein; EBA – erythrocyte binding antigen; SERA – serine repeat antigen; GLURP – glutamate rich protein; LSA – liver stage antigen; CSP – circumsporozoite protein; TRAP thrombospondin-related adhesive protein; AARP-1 - asparagine and aspartate rich protein-1; ChAd63 - chimpanzee adenovirus 63; MVA – modified vaccinia virus Ankara; GIA – growth inhibition assay; IgG – immunoglobulin gamma; C-

terminal – carboxy terminal; FVO - falciparum Vietnam oak-knoll; PNG – Papua New Guinea. * Cochrane review of ten trials of SPf66 concluded that vaccine had no significant efficacy (apart from in South America) and that there was no justification for further trials (Graves and Gelband, 2006).

Chapter 2 - Calling *msp1* block 2 alleles from short read data

2.1 Introduction

MSP-1 is encoded by a 5 kb gene (*msp1*) on chromosome 9. Early studies showed that this gene was composed of conserved and polymorphic regions (named blocks 1-17) (Tanabe et al., 1987). The most polymorphic region of the gene is block 2 (Miller et al., 1993). The polymorphism at block 2 has been classified into three main allelic types based on homology with laboratory lines: K1-like, MAD20-like and RO-33-like (Miller et al., 1993). K1-like and MAD20-like *msp1* block 2 alleles contain highly polymorphic repeat sequences that are flanked at the 5' and 3' ends by non-repeat sequences. Both K1-like and MAD20-like alleles encode repeating tripeptide motifs. These motifs and the non-repeat flanking sequence are distinct between the allelic families. Whilst the non-repetitive flanking sequences are conserved between members of the same allelic family, expansion, contraction and sequence variation of the repeats in both K1-like and MAD20-like families results in the presence of multiple individual alleles within each family. Analysis of polymerase chain reaction (PCR) fragment sizes shows around twenty distinct lengths present in each of the K1-like and MAD20-like allele families among large numbers of studies (Branch et al., 2001, Takala et al., 2006). Sequencing of these PCR products shows that variation in the repeat structure results in almost four times more K1-like alleles and almost twice as many MAD20-like alleles (Noranate et al., 2009). The RO-33-like allele does not contain a repeat sequence and is conserved with only six sites in which amino acid substitutions have been identified in total (including the results of this study, see below section 2.3.9), generating seven different alleles (Noranate et al., 2009, Tanabe et al., 2013). MR recombinant alleles, so named due to apparent recombination between the MAD20-like and RO-33-like alleles, resulting in a sequence with homology at the 5' end to MAD20-like alleles and at the 3' end to RO-33-like alleles, were originally identified in East Africa (Takala et al., 2002) and subsequently in West Africa, Asia and South American (Takala et al., 2006, Noranate et al., 2009, Tanabe et al., 2007b).

Almost seventy studies (for list see appendix 7.1) have assayed the *msp1* block 2 genotypes found in parasite populations using PCR to specifically amplify each of the three main allelic families (Viryakosol et al., 1995). Genotyping of the *msp1* block 2 locus by this method has shown that MAD20- and K1-like alleles are the most common, each representing about two fifths of all alleles with RO-33 like making up the remaining fifth (table 2.1). In parasites sampled from Africa, K1-like alleles dominate, whereas in Asian parasites MAD20-like alleles have a higher frequency (Table 2.1) (Conway et al., 2000). In South America MAD20-like alleles are the most common overall (accounting for four fifths of all alleles) but there is a large degree of variation in allele frequencies between study sites (Silva et al., 2000), consistent with a higher degree of isolation of parasite populations.

Continent	Region	K1-like	MAD20-like	RO-33-like
Africa	West Africa	1832 (42.4)	1241 (28.7)	1248 (28.9)
	East Africa	2127 (42.1)	1453 (28.7)	1470 (29.1)
	central Africa	539 (37.7)	417 (29.2)	473 (33.1)
	Southern Africa	43 (58.9)	20 (27.4)	10 (13.7)
	North Africa	93 (30.2)	108 (35.1)	107 (34.7)
	Total	4634 (41.4)	3239 (29.0)	3308 (29.6)
Asia	South East Asia	728 (34.7)	1005 (47.9)	367 (17.5)
	South Asia	413 (31.1)	488 (36.8)	426 (32.1)
	West Asia	33 (41.2)	16 (20.0)	31 (38.8)
	Total	1174 (33.5)	1509 (43.0)	824 (23.5)
Oceania	Melanesia	95 (19.8)	166 (34.7)	218 (45.5)
South America	Amazon basin	195 (11.3)	1386 (80.0)	151 (8.72)
Total		6256 (36.2)	6379 (36.9)	4637 (26.8)

Table 2.1 Distribution of *m*sp1 block 2 allelic families by region from published studies. Literature searches were performed to find all studies using *m*sp1 genotyping. Total counts, divided by region, are shown with percentage of total in parenthesis. The full list of studies can be found in appendix 7.1.

Whilst allele frequencies vary between populations, the maintenance of multiple alleles at the *msp1* block 2 locus across populations suggests that natural selection is operating to keep these alleles in the population (Conway et al., 2000). Host immunological memory is predicted to select for the presence of multiple alleles, indicating that *msp1* block 2 is an immune target (Weedall and Conway, 2010). Indeed, antibodies recognising *msp1* block 2, in particular IgG3 against the polymorphic repeat sequences, have been shown to correlate with protection from malarial disease in several West African populations (Cavanagh et al., 2004, Conway et al., 2000, Polley, 2003 #50). Although this finding was not replicated in all studies (Gray et al., 2007, Osier et al., 2008), a meta-analysis found an association between K1-like antibodies and protection from malaria (Fowkes et al., 2010). Furthermore, vaccination with an antigen based on *msp1* block 2 resulted in protection of two out of four *Aotus lemurinus griseamembra* monkeys from developing high parasitaemia after challenge with the virulent FVO parasite strain (Cavanagh et al., 2014). Human antibodies have been shown to inhibit parasite growth in the presence of monocytes (Galamo et al., 2009) and rabbit antibodies have been found to inhibit merozoite invasion with the addition of active complement (Osier et al., 2014a), suggesting that antibodies against MSP-1 block 2 function via secondary mechanisms rather than direct blocking of RBC invasion.

Antibodies against MSP-1 block 2 antigens have been repeatedly found to be against polymorphic epitopes, that are either major allele type specific or that reflect sub-typic polymorphism (Polley et al., 2003b, Cavanagh et al., 2004, Cavanagh et al., 1998, Cavanagh and McBride, 1997, Conway et al., 2000, Mawili-Mboumba et al., 2003, Ekala et al., 2002, Jouin et al., 2005, Jouin et al., 2001, Kimbi et al., 2004, Da Silveira et al., 1999, Scopel et al., 2005). Surveillance of allele frequencies in parasite populations is essential in both determining the formulation of polymorphic vaccines and monitoring their impact to identify possible vaccine escape (Barry and Arnott, 2014, Ouattara et al., 2015, Takala and Plowe, 2009). Whilst the development of high-throughput technologies for genotyping relevant

parasite loci will need to be developed in order to perform surveillance, as has been done for drug resistance loci (Carnevale et al., 2007), large scale whole genome sequencing projects provide a rich source of parasite genetic data (consortium, 2015).

Due to the interest in *msh1*, over thirty studies have sequenced this gene or the block 2 fragment over the past 30 years (for a list of all studies see appendix 7.2). All but one of these studies have used chain termination sequencing, which produces long (up to 600 bp) sequence reads that span the whole of block 2 (Sanger and Coulson, 1975, Sanger et al., 1977). One study used pyro-sequencing, which produces sequencing reads of up to 300 bp (Juliano et al., 2010), long enough to capture the vast majority of *msh1* block 2 sequences. These long read sequences provide a rich archive of *msh1* block 2 sequence data.

The capacity to perform massively parallel sequencing using the Illumina sequencing platform (Meyer and Kircher, 2010) has allowed for the generation of thousands of whole *P. falciparum* genomes (Miotto et al., 2015). The Pf3k project (<https://www.malariagen.net/projects/pf3k>) is a collaboration that aims to collate sequence data for at least three thousand *P. falciparum* parasite isolates from across the global distribution of the species (consortium, 2015). The project currently has short read sequence data from 2400 parasites isolates from 14 countries of Africa and Asia (consortium, 2015) and represents the largest resource of *P. falciparum* genetic data currently available.

Processing of short read sequence data typically involves alignment of reads to a reference sequence. Single nucleotide polymorphisms (SNPs) can then be determined as the positions where the aligned reads differ from the reference sequence (MacLean et al., 2009). This approach for cataloguing genetic variation cannot be applied when there is an extended polymorphic sequence as only reads with the same polymorphism as the reference sequence will align (MacLean et al., 2009). Additionally, alignment based SNP calling can be confounded by the presence of repetitive sequence, as the alignment algorithm cannot place a read in a region where there are multiple

possible alignments (Li and Durbin, 2009). The *msp1* block 2 locus contains both extended polymorphism and repeat sequence (Miller et al., 1993, Tanabe et al., 1987) and therefore *msp1* block 2 polymorphism cannot be assessed reliably from short read data by alignment to the *P. falciparum* reference sequence. In fact this and other repetitive and highly polymorphic loci are normally removed from analysis of genome wide polymorphism (Amambua-Ngwa et al., 2012a).

One method that can circumvent the presence of highly polymorphic sequence data is the use of algorithms to assemble the sequence reads without the use of a reference sequence. This approach, known as *de novo* assembly, standardly uses one of two types of algorithm. The first assemblers used overlap-layout-consensus (OLC) algorithms which find the best overlaps between reads and finds a path that goes through each read just once (Hamiltonian path) before joining overlapping reads to assemble contigs (Staden, 1979) figure 2.1). The second generation of assemblers use De Bruijn graph (DBG) based algorithms which first split reads into sub-strings of length k (with the maximum value of one less than the read length), known as k -mers, before constructing a graph with nodes for each k -mer and edges for the overlap between each k -mer. DBG based algorithms then determine contigs by finding a path using all edges of this graph (Eulerian path) that links the greatest number of k -mers (Idury and Waterman, 1995) figure 2.1). DBG assemblers have greater efficiency than OLC assemblers, enabling their application to the large datasets such as those generated by next generation whole genome sequencing platforms meaning that this algorithm has now been widely adopted (Li et al., 2012). Both assembly algorithms are not able to resolve repeats longer than the k -mer or read length as these result in the formation of bubbles and branches in the graph (Miller et al., 2010) figure 2.2 a); the only way to overcome this issue is by increasing the k -mer or read length (Li et al., 2012) until it is longer than the repeat region (figure 2.2 b).

The coverage depth at any position in the genome is defined as the number of reads containing the base at that position. The average coverage depth is the mean of coverage depth of all bases which is equal to the number of reads multiplied by the read length and divided by the length of the input

sequence. Increasing coverage depth will aid *de novo* assembly as a greater amount of information is provided by the increased number of reads (Li et al., 2012). Increasing coverage depth will also aid the error correction functions present in many algorithms that aim to correct sequencing errors by providing a greater number of reads or *k-mers* that can be used to determine the correct path through a branched graph (Miller et al., 2010). When the read or *k-mer* length is only marginally longer than a repeat sequence, increased coverage depth will aid *de novo* assemblers to resolve the sequence by stochastically increasing the number of reads or *k-mers* that span the repeat.

Many DBG assemblers are available but only several, AllPaths-LG (Gnerre et al., 2011) SOAPdenovo (Li et al., 2010) and Velvet (Zerbino and Birney, 2008), allow long (> 31) *k-mer* lengths (Li et al., 2012). Comparison of the three packages showed similar performance when assembling eukaryotic genomes (Zhang et al., 2011). Velvet was chosen for use in this project due to its common use in assembly of Illumina data, its ability to handle long *k-mer* lengths and its ease of use in comparison to other packages. Velvet also employs algorithms that remove sequencing errors through analysis of *k-mer* coverage and the DBG structure, leading to highly reliable sequence data (Zerbino and Birney, 2008).

Another approach to calling highly polymorphic genomic regions is the construction of sequence libraries that contain a catalogue of the known diversity to which sequences can be aligned. Such an approach has been used for calling sequence polymorphisms at the HLA locus, the most polymorphic site in the human genome (Robinson et al., 2001). The major drawback of this approach is that it yields only the allelic type and not the complete sequence. Given the diversity within the allelic types of *msp1* block 2 (Noranate et al., 2009) and other polymorphic repeat sequences (Anders et al., 1988, Anders et al., 1993) this results in losing a large degree of information. Furthermore, the presence of novel alleles resulting from recombination between two different allelic types, as has been reported for *msp1* block 2 (Takala et al., 2002), will be missed by this approach as such alleles

will be called as mixed genotype infections (infections in which the host contains parasites with two or more genotypes as a result of superinfection).

We aim to employ both *de novo* assembly and alignment to a library of sequences to explore *msp1* block 2 polymorphism in the Pf3k short read data. The use of short read data in the study of *msp1* block 2 polymorphism can contribute to vaccine candidate design as well as surveillance of allele frequencies. Furthermore, this hybrid approach can also be applied to study other *P. falciparum* vaccine candidates, many of which are highly polymorphic and contain repeat sequences. The results of optimising and using both these approaches for *msp1* block 2 are presented below.

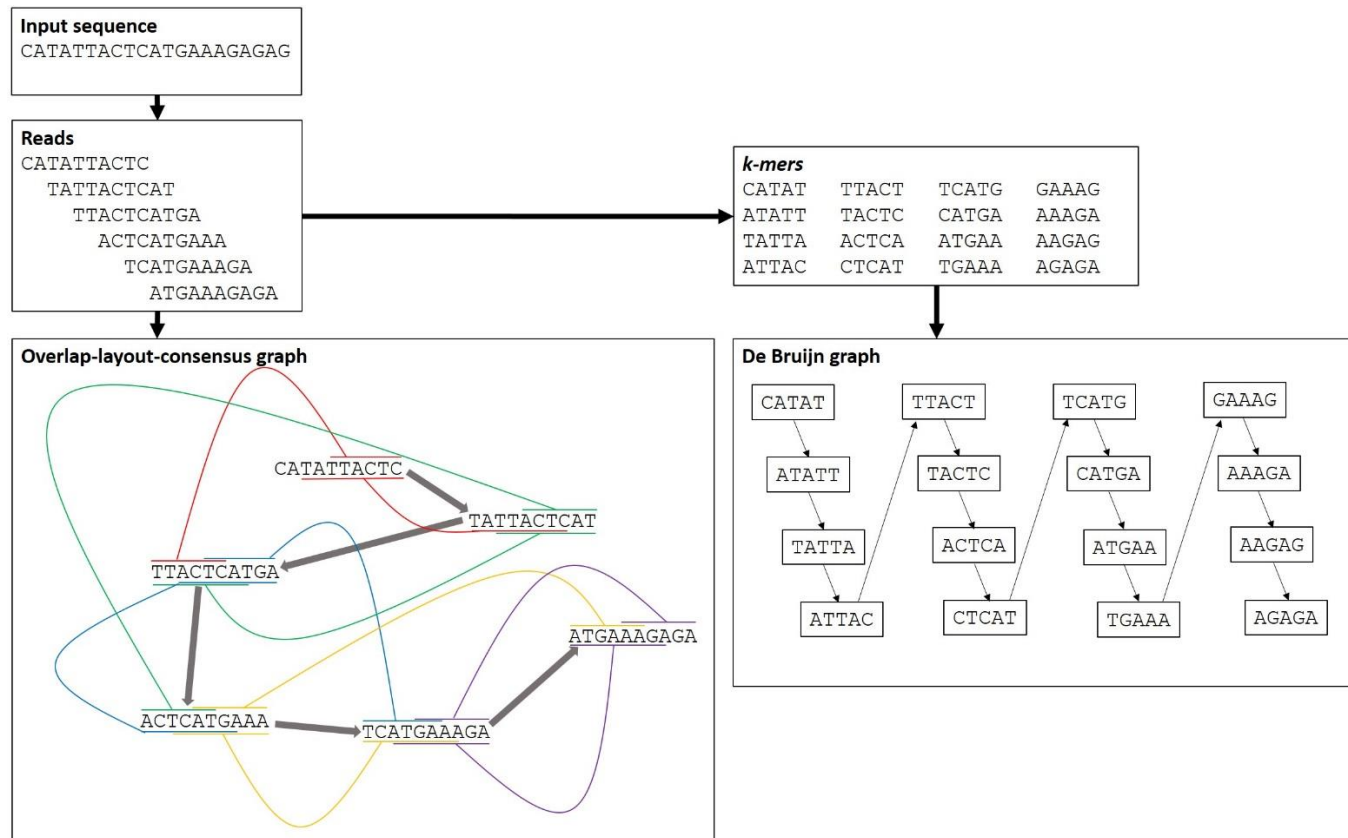
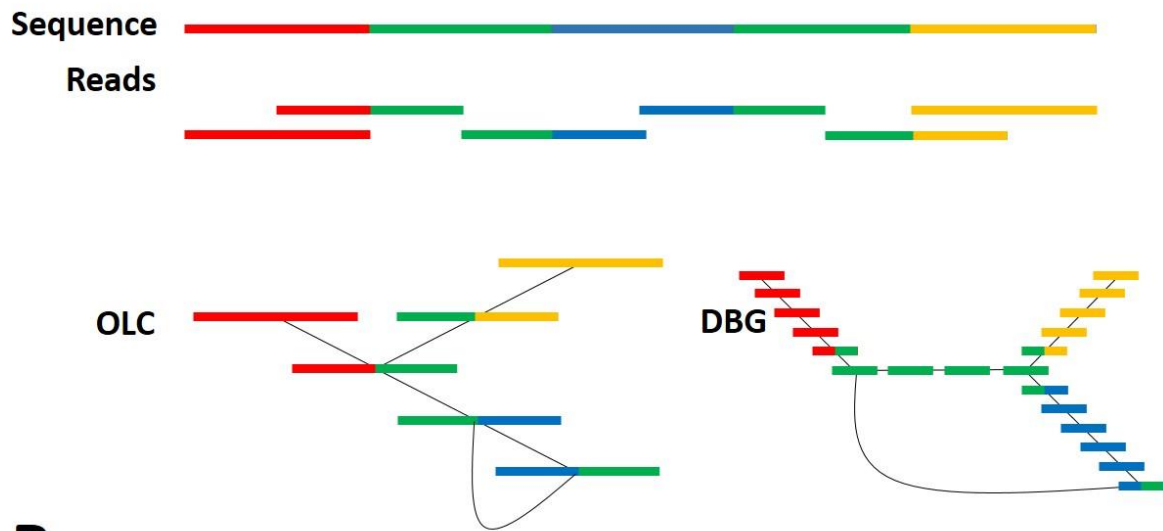


Figure 2.1 Schematic representation of two main assembly algorithms. In this diagram, an input sequence generates 6 reads of 10 bp length (top right). Overlap-layout-consensus (OLC) algorithms will find the overlaps longer than a cut-off (5 bp is used here) between these reads and join them in a graph (bottom left) with each read as a node and overlaps as edges (coloured lines); note that there are up to three edges between reads. The contig sequence will be determined by finding a Hamiltonian path (grey arrows) between the reads. De Bruijn graph (DBG) algorithms will first split the reads into all possible sub-strings (*k-mers*) of length *k* (here *k* is 5) before building a graph with *k-mers* as nodes and the edges as the overlap of length *k*-1 between *k-mers*. The Euler path between these nodes then determines the contig; note that there is just one edge between each *k-mer*.

A



B

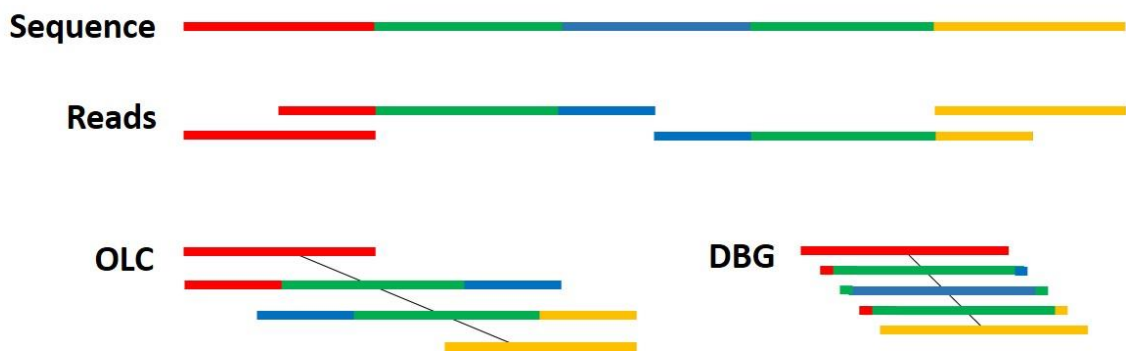


Figure 2.2 Schematic representation of how assembly algorithms represent repeat sequences. An input sequence has a repeated sequence (green) separated by three unique sequences (red, blue and yellow). When read and k -mer length are shorter than the repeat (A) all repeat containing reads will be present in the graph generated by overlap-layout-consensus (OLC) algorithms whereas the k -mers containing the repeat sequence will be collapsed into a single sequence in the De Bruijn graph (DBG). Both graphs contain branches and loops (known as bubbles) as a result of the repeat sequence which cannot be resolved. When the read or k -mer length spans the repeat sequence (B) both algorithms can use the unique sequence at either end of the repeated sequence to produce a single path through the graph and reconstruct the original sequence accurately. Figure adapted from (Li et al., 2012)

2.2 Materials and methods

2.2.1 Long read sequence data

Long read sequence data (LRD) was downloaded from GenBank (Benson et al., 2013). GenBank was searched with the search terms: “*plasmodium falciparum* [organism] *msp1*”; “*plasmodium falciparum* [organism] *msa1*”; and “*plasmodium falciparum* [organism] *gp195*” on 4th December 2015¹. All sequence 1831 results were downloaded and curated for presence of a complete *msp1* block 2 sequence, found in 1007 of these. Removal of sequences from identical laboratory strains resulted in a total of 964 sequences (381 K1-like, 350 MAD20-like, 202 RO-33-like and 31 MR-like). The list of studies and accession numbers for all sequences can be found in appendix 7.2 and a full list of sequences can be found in additional data file “long_read_sequences.fa”.

2.2.2 Generation of synthetic reads from long read sequence data

In order to generate short read data for known *msp1* block 2 sequences for use in validating novel bioinformatic approaches, short reads were created *in silico* from *msp1* block 2 sequences obtained by long read sequencing and deposited in GenBank. The *msp1* block 2 sequence from each of the 964 LRD sequences (section 2.2.1) was inserted into version 3 of the *P. falciparum* 3D7 reference sequence for the *msp1* gene and 2kb of sequence upstream of the start of the gene (chr9:1199812-1206974) downloaded from PlasmoDB (Aurrecoechea et al., 2009). The python script to_perfect_reads, part of the package Fastaq (downloaded from <https://github.com/sanger-pathogens/Fastaq>) was modified to add quality scores from fastq files downloaded from Pf3k project (appendix 7.3) and used to create synthetic reads for each of the 964 *msp1* block 2 sequences with flanking regions from the 3D7 reference sequence (figure 2.3). The majority of samples in the Pf3k dataset were sequenced with 100 or 75 bp read lengths. Therefore synthetic reads were created at both 100 bp and 75 bp length with a mean insert size and standard deviation (SD) representative of

¹ GenBank search ignores the hyphen hence “*msp-1*” and “*msa-1*” are effectively included in these searches

the Pf3k data set (mean insert size² of 250 bp, SD 83 bp for 100 bp reads and mean insert size of 277 bp, SD 83 bp for 75 bp reads).

2.2.3 Illumina paired-end short read sequence data

Binary alignment/map (BAM) files and metadata were downloaded from Pf3k release version 4.0 (available at ftp://ngs.sanger.ac.uk/production/pf3k/release_4/). For 113 of the 2518 Pf3k samples two sequencing runs had been performed; in these cases the run with the highest mean coverage was kept and the other discarded. The resulting dataset contains sequencing reads from 2400 isolates collected by 10 studies from 26 sites in 15 countries with an additional 5 laboratory line sequences, which were not included in further analysis. The read length ranged from 30-100 bp, with a genome-wide mean coverage ranging from 1- to 676-fold. All samples, including those with low genome-wide coverage were included as the coverage of *msp1* block 2 cannot be determined from this summary data and may well be higher due to increase GC content relative to non-genic regions.

2.2.4 De novo assembly

In an order to capture reads originating from *msp1* block 2 sequences that are not mapped because of sequence polymorphism (see above section 2.1) read pairs with at least one mate mapping to the region of chromosome 9 encompassing *msp1* and 2kb upstream (Pf3D7_09_v3:199812 – 1206974) were extracted from Pf3k BAM files with SAMtools (Li et al., 2009). This region was chosen as *msp1* block 2 sits toward the 5' end of a ~5 kb gene and the vast majority of insert sizes are under 2 kb hence the vast majority of mate pairs of reads originating from *msp1* block 2 will map within this region. Reads were also extracted from BAM files produced from alignment to an *msp1* block 2

² The insert size is the size of the DNA fragment from which both mates of a pair of sequence reads are created during Illumina sequencing.

sequence library. *De novo* assembly of these reads was performed with Velvet (version 1.2.10) (Zerbino and Birney, 2008), with a *k-mer* length of 81 (see section 2.3.2).

2.2.5 Alignment of short reads

Raw reads were extracted from BAM files into Fastq files using SAMtools (version 1.3) (Li et al., 2009). These reads were aligned to a reference library of 15 *msh1* block 2 sequences (section 2.3.4, appendix 7.4) using BWA-MEM (version 0.7.5a-r405) (Li, 2013) with default parameters. The resultant sequence alignment/map (SAM) files were sorted, indexed and compressed using Sambamba (version 0.6.0) (Tarasov et al., 2015). SAMtools (Li et al., 2009) was used to get the alignment statistics and thus determine the coverage over each sequence of the library. Coverage was calculated for each allelic type as the number of bases in reads align to reference sequences of that allelic type divided by the total length of reference sequences of that allelic type. Coverage was calculated for each sample by summing the coverage of each allelic type.

2.2.6 Data analysis

Data was analysed using the statistical analysis tool R (Team, 2008) with additional package ggplot2 (Wickham, 2009) for graphical functions. Sequences were aligned using MAFFT (Katoh et al., 2002).

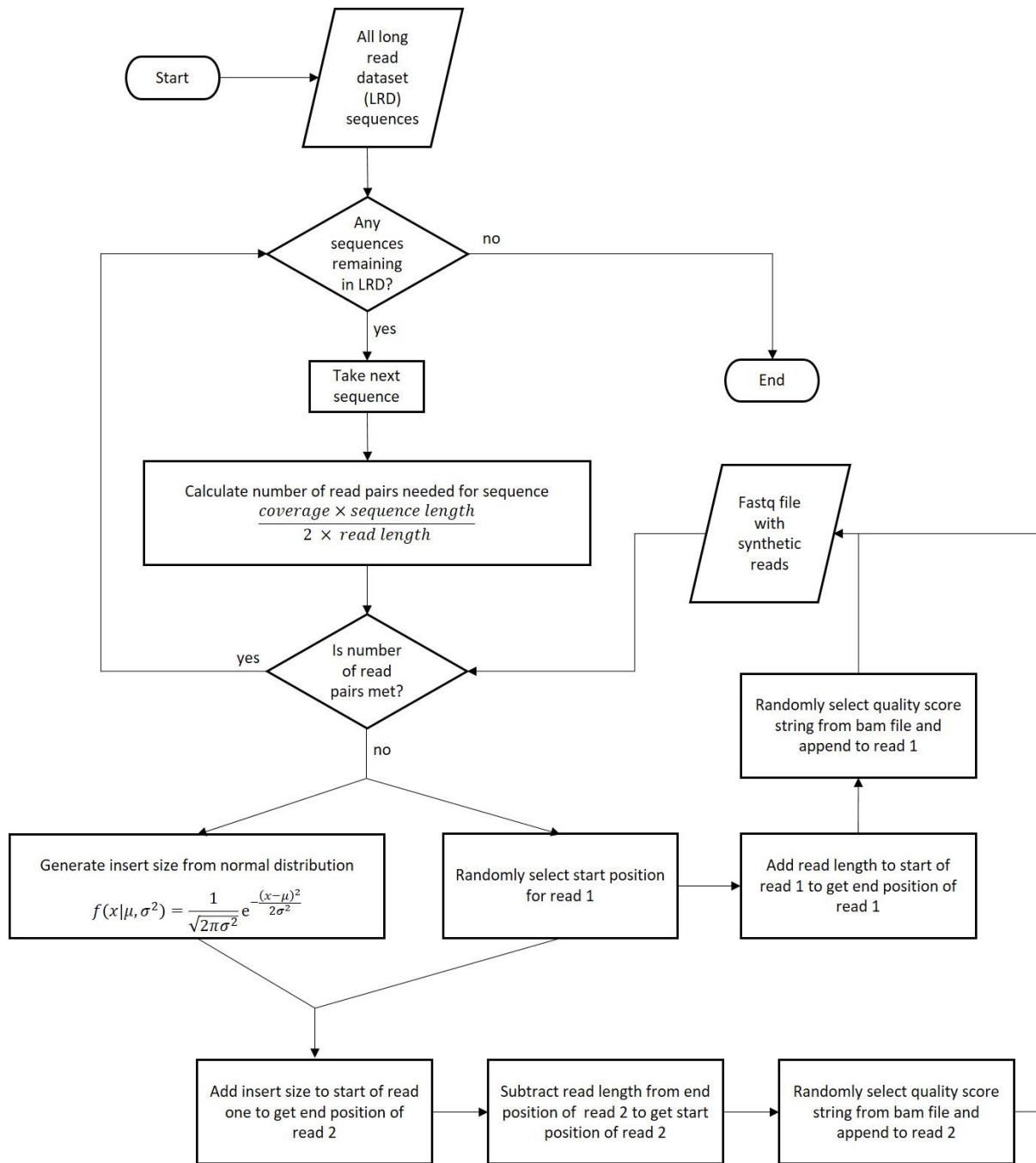


Figure 2.3 Flow chart showing how synthetic reads were generated from long read sequence data. 964 long read MSP-1 block 2 sequences were extracted from GenBank (see section 2.2.1). For a given read length and coverage depth, synthetic reads were generated for each sequence. Read pairs were created by to_perfect_reads (downloaded from <https://github.com/sanger-pathogens/Fastaq>; appendix 7.3); for each read pair the start position for read one was randomly selected. The insert size was selected from a binomial distribution of insert sizes around the mean insert size and used to determine the end position of read two. Quality scores for both reads were randomly selected from a bam file from the Pf3k project and added to the fasta file along with the read pair.

2.3 Results

2.3.1 Creation of long read sequence dataset for *msh1* block 2

Multiple studies have used long read sequencing methods to sequence the block 2 region of *msh1* block 2. In order to build a database for the validation and benchmarking of novel methods developed for analysis of *msh1* block 2 short read sequence data, long read sequences deposited in GenBank were downloaded. From the 1831 results of GenBank searches, removal of sequences for identical lab isolates, spurious results, sequences lacking a complete block 2 sequence resulted in 964 sequences. The *msh1* block 2 sequence was extracted from each of the 964 sequences and entered into the dataset, known from here on as the long read dataset (LRD). Thirty five of the 36 studies contributing sequences to LRD used Sanger sequencing, one study used pyrosequencing (Juliano et al., 2010) (for a list of accession numbers and studies see appendix 7.2).

2.3.2 *De novo* assembly optimised for reconstruction of *msh1* block 2 sequences

Due to the highly polymorphic nature of the *msh1* block 2 locus, aligning short read data to a reference sequence would fail as reads from non-reference-like alleles could not be mapped (MacLean et al., 2009). *De novo* assembly of reads could avoid this problem as reads are assembled by finding the sequence overlaps between reads so homology with a reference sequence is not required. Velvet is a collection of algorithms that can assemble short read data using De Bruijn graphs (Zerbino and Birney, 2008). The most important parameter to optimise for these algorithms is the length (k) of sub-sequences (termed k -mers) that the reads are broken into before construction of the De Bruijn graph (Zerbino, 2010). This is because a longer k -mer length will facilitate assembly of repeat regions by bridging the repeat (figure 2.4). However, longer k -mer lengths will result in a lower k -mer depth (as fewer k -mers can be made from the reads), resulting in

a lower probability of assembly in an analogous fashion to the coverage depth (figure 2.5) (Li et al., 2012).

Assembly of synthetic short read data created *in silico* from LRD sequences (section 2.2.2) was used to optimise the *k-mer* length for Velvet and assess performance. Testing over a range of *k-mer* lengths between 31 and 99, demonstrated that a *k-mer* length of 81 was optimal for assembly of *msp1* block 2 sequences of all allelic types (figure 2.4). Usage of Velvet with this *k-mer* length resulted in correct assembly of 93.6% (902/964) of *msp1* block 2 sequences. The fact that a high proportion of *msp1* block 2 sequences could be assembled from synthetic short read data, and that no assemblies contained errors was encouraging. However, the fact that a lower proportion of K1-like sequences was assembled compared to MAD20-like and RO-33-like, warranted concern that use of *de novo* assembly to capture *msp1* block 2 sequences would lead to an allele bias.

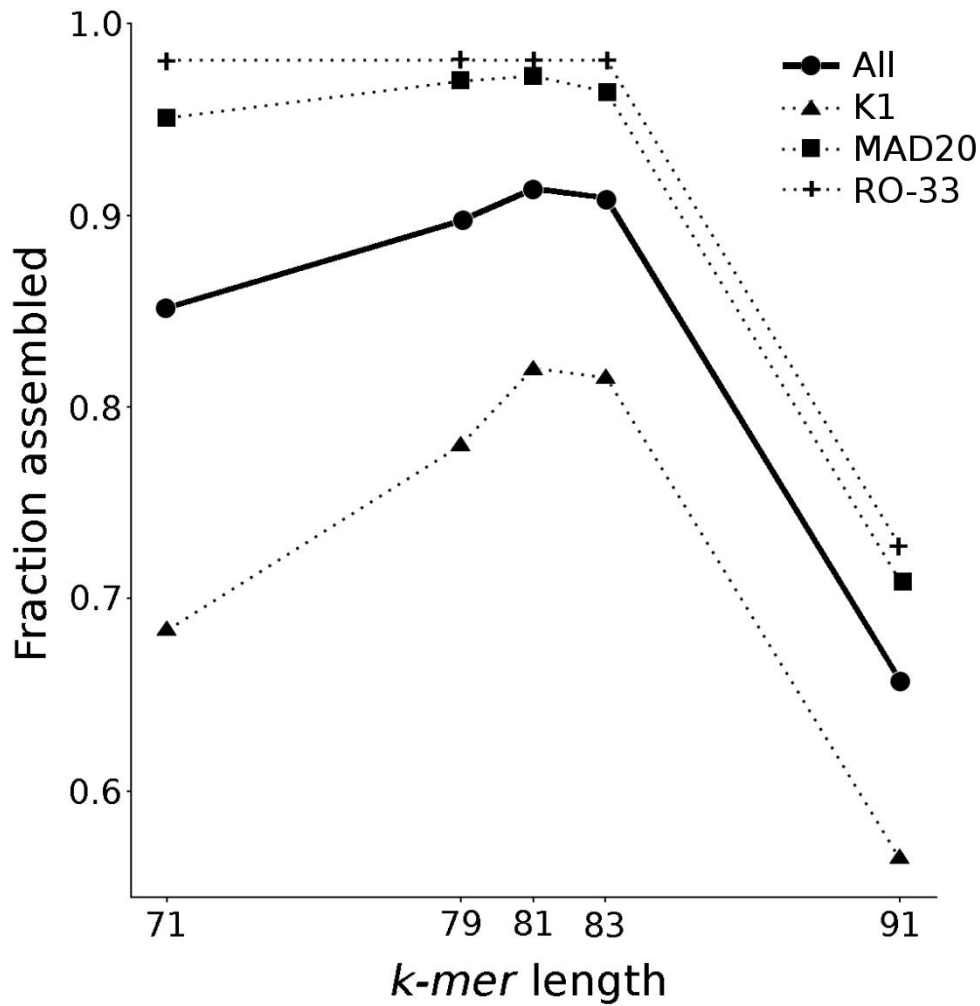


Figure 2.4. The effect of *k-mer* length on the fraction of *msp1* block 2 sequences assembled by Velvet. 964 *msp1* block 2 sequences from the LRD were used to create synthetic reads of 100 bp in length with a coverage of 50-fold. These reads were assembled using Velvet (Zerbino and Birney, 2008) with a range of *k-mer* lengths. The resulting contigs were then scanned for the presence of the correct *msp1* block 2 sequence. The fraction of *msp1* block 2 sequences that were fully assembled for all 964 sequences is shown (solid line). The fraction of sequences fully assembled for K1-like sequences (triangles, $n = 392$), MAD20-like sequences (squares, $n = 354$) and RO-33-like sequences (crosses, $n = 204$) are also shown (dotted lines).

2.3.3 Optimised *de novo* assembly of *msp1* block 2 sequences results in bias towards short alleles

The algorithms used by Velvet and other *de novo* assemblers to assemble short read data are unable to resolve a perfect repeat sequence that is longer than the *k-mer* length (Li et al., 2012). As the majority of the *msp1* block 2 sequence is repetitive, with expanded repeats resulting in longer sequences, it was necessary to investigate the effect of sequence length on the likelihood of assembly by Velvet with an optimised *k-mer* length of 81. It was found that synthetic reads created from longer sequences were less likely to be assembled by Velvet (figure 2.4, $p < 0.001$, Wilcoxon signed rank test). This is of concern as *de novo* assembly of *msp1* block 2 sequences will produce a bias towards shorter sequences. When using reads shorter than 82 bp (which includes the 75 bp read length used for sequencing 615 isolates in the Pf3k project), *de novo* assembly has to be performed with a sub-optimal *k-mer* length of 74. This will further decrease the possibility of assembly of longer repeat sequences as the *k-mers* need to be long enough to span the repeat region in order to allow resolution by the assembly algorithm (figure 2.2). For this reason, only Pf3k samples that had been sequenced with a read length > 82 bp were used for *de novo* assembly.

To ascertain the effect of coverage depth on the assembly of long sequences, synthetic reads were generated at a range of coverage depths (i.e. with a range of total number of reads) from the *msp1* block sequence of lab isolate Palo Alto (270 bp)(Chang et al., 1988). The synthetic read generation algorithm will position reads randomly (section 2.2.2) and, as the position of the reads will influence the ability to assemble them (see section 2.1), reads were generated 10 times for each coverage depth assayed. Velvet was then used to assemble the reads and the resulting contigs were scanned for the complete Palo Alto *msp1* block 2 sequence. As expected, increasing coverage depth improves the chances of complete assembly of the *msp1* block 2 region (figure 2.5, $p = 0.96$, $p < 0.001$). At a coverage depth of over 80-fold, assembly of this long block 2 sequence is consistently achieved.

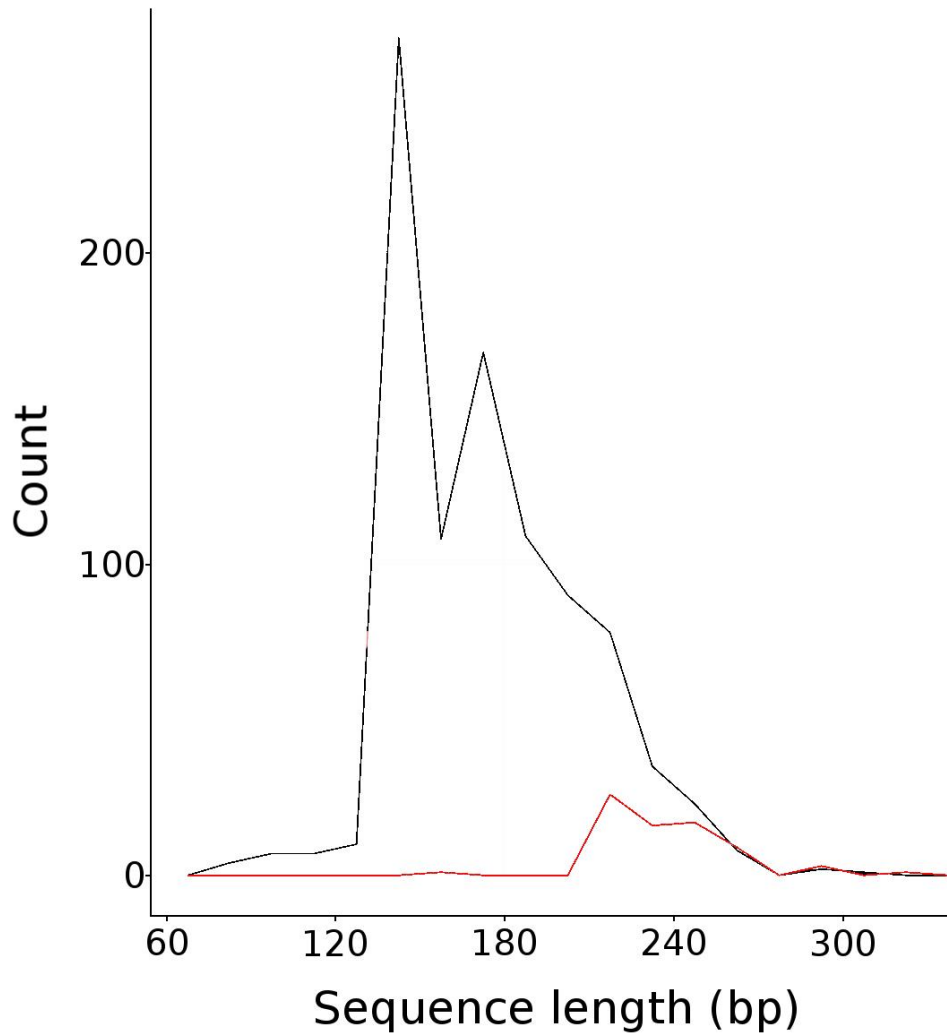


Figure 2.5. Frequency distributions of length of block two sequence for assembled and unassembled sequences show bias towards assembly of shorter sequences. Synthetic reads created from 964 *msp1* block 2 sequences were assembled using Velvet (Zerbino and Birney, 2008). The frequency distributions of the lengths of the original *msp1* block 2 sequences are shown for successfully assembled (black line, n = 902) and unassembled (red line, n = 62) sequences are shown (sequences lengths are grouped with a bin width of 5 bp).

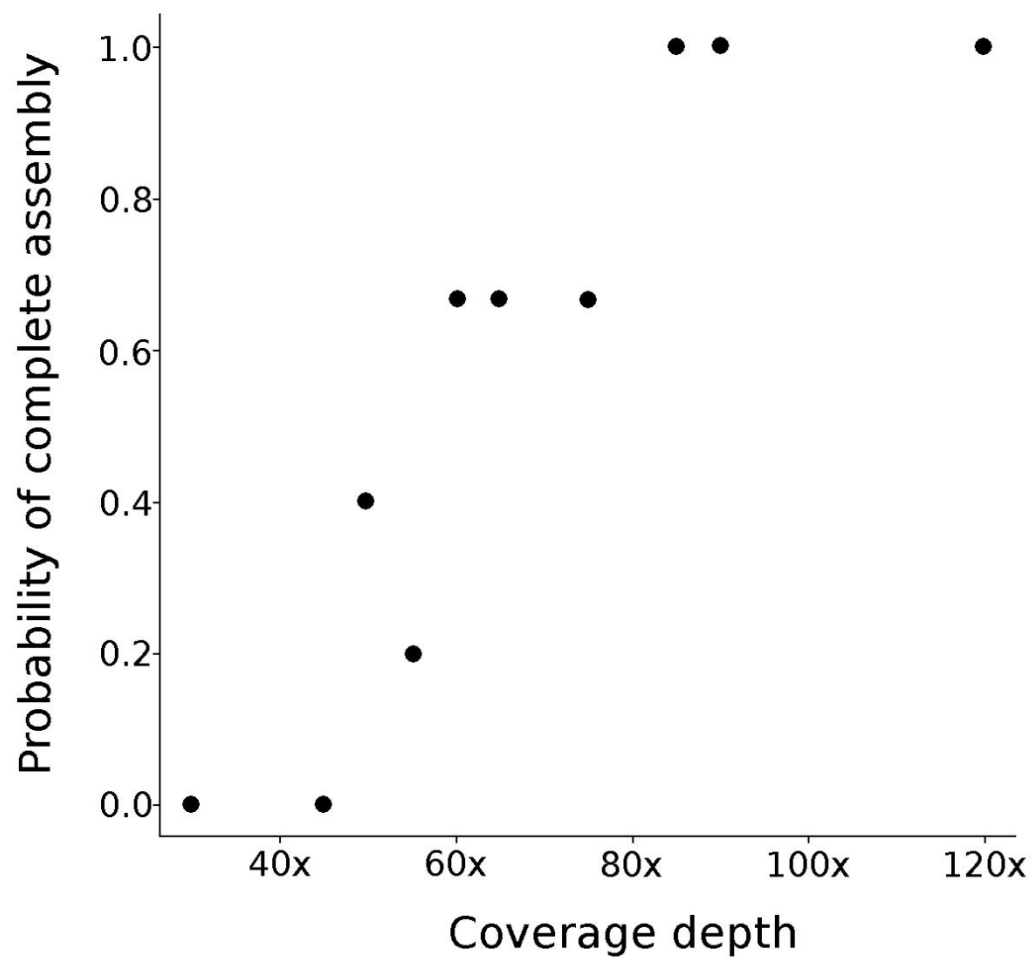


Figure 2.6 Probability of complete assembly of *msp1* block 2 is dependent on depth of coverage.

Synthetic reads were generated *in silico* from the Palo Alto sequence of *msp1* at 10 different coverage depths. Reads were generated 10 times for each coverage depth and then assembled using Velvet (Zerbino and Birney, 2008) and the presence of the complete block 2 sequence was determined. The probability of assembling the whole *msp1* block 2 sequence is shown. There is a strong and significant correlation between coverage depth and the probability of complete assembly of the Palo Alto *msp1* block 2 sequence ($\rho = 0.96$, $p < 0.001$).

2.3.4 Creation of a reference library of *msh1* block 2 sequences allows for reads to be aligned

In order to overcome the bias in *de novo* assembly of *msh1* block 2 short read data and allow allele calls to be made from 75 bp sequence reads, a library of *msh1* block 2 sequences was created to be used as reference sequences for alignment of short read data. To create the library, all LRD sequences were grouped into allelic type (MR-like recombinant alleles were excluded). Sequences were aligned with all other sequences in the allelic group and the sequence closest to the consensus sequence (i.e. most similar to all sequences) was then added to the library. This generated the first library containing one sequence per allelic type (three sequences in total). Synthetic reads were then aligned to the library using the basic BWA algorithm, which does not allow reads to be gapped (Li and Durbin, 2009). The number of reads mapping for each sequence was analysed and the sequences for which fewest reads were mapped were then aligned in their allelic groups. Again the sequence closest to the consensus sequence was added to the library to give the second library containing two sequences per allelic type (6 in total). This process was repeated iteratively until 10 libraries were generated with one to 10 sequences per allelic type (three to 30 sequences in total), see figure 2.7.

The 10 libraries were then tested by aligning the same sets of 100 bp and 75 bp synthetic reads generated from all LRD sequences to each library with BWA-MEM, which tolerates gaps in the alignment (Li, 2013). Use of the library containing 5 *msh1* block 2 sequences per allelic type (15 sequences total) resulted in the alignment of the greatest number of 75 and 100 bp reads (figure 2.8). Aligning to this library also resulted in at least 67 reads (approximately 35-fold coverage) mapping for each sequence. The sequences used in this library are listed in appendix 7.4. This library will be used for all subsequent alignment and will be referred to as the *msh1*-1 block 2 reference library (*msh1b2RefLib*).

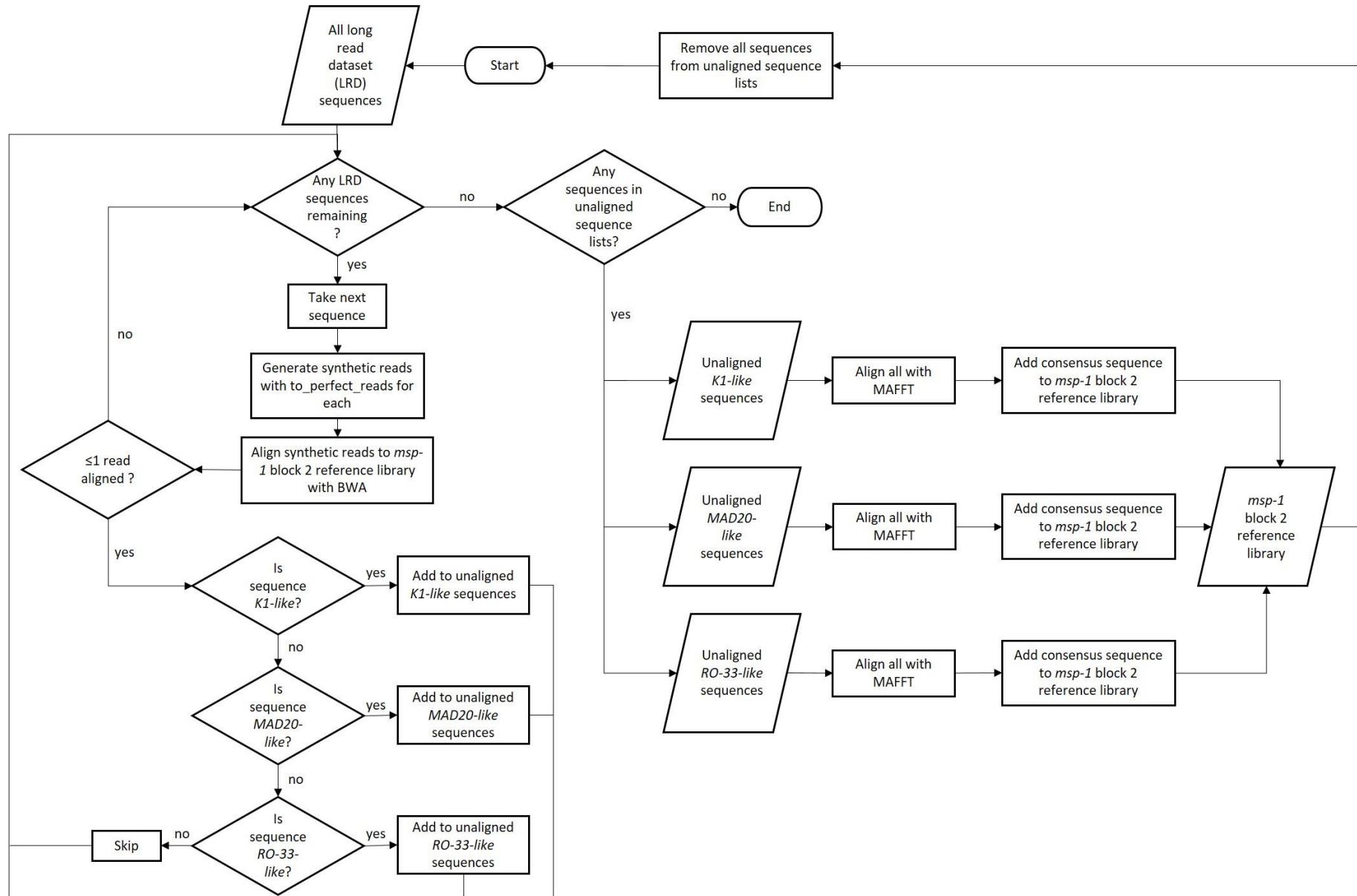


Figure 2.7 Flow chart showing method for creaion of *msh-1* block 2 reference libraray.

All *msh-1* block 2 sequences were grouped by allelic type and aligned. The sequence closest to the consensus from each type was used as the starting sequence for the reference library. Synthetic reads were then generated from all sequences in the LRD and aligned to the reference library with BWA. If just one or no reads aligned to the reference library for a given sequence, the allelic type of that sequence was determined and the sequence was added to a list of unaligned sequences of the same type. Once all the sequences in the LRD had been put through this process all the sequences of each allelic type in the list of unaligned sequences were aligned with each other using MAFFT. The sequence closest to the consensus sequence was then added to the reference library. After each iteration, the library was saved so that it could be tested for its ability to align to reads from the LRD using the algorithm BWA-MEM (figure 2.8). The protocol was run for nine iterations to produce 10 reference libraries (the first reference library containing the original three sequences and the tenth containing 30 sequences, 10 for each allelic type).

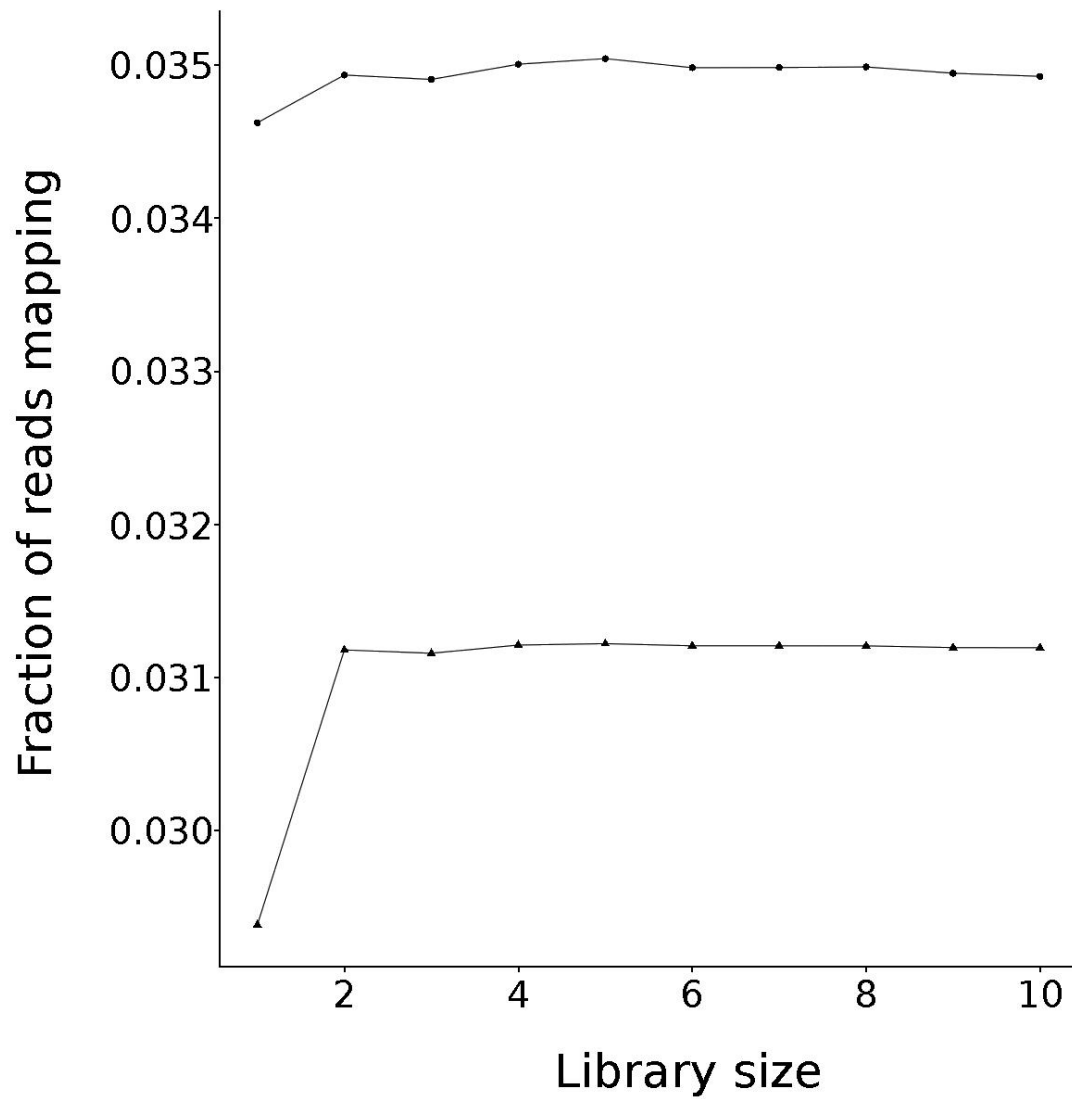


Figure 2.8 Effect of the number of sequences in the reference library on the number of reads mapped. Reference libraries were created with between one and 10 sequences per major allelic type (3 to 30 sequences total). Synthetic reads created from 964 long read sequences (section 2.2.2) and were aligned to the sequence libraries using BWA-MEM (Li, 2013). The fraction of all reads that were mapped to the reference library is shown for read lengths of 100 bp (circles) and 75 bp (triangles).

2.3.5 Alignment to library of *msp1* block 2 sequences enables unbiased calling of allelic type from synthetic short read data

It was expected that alignment to the *msp1* block 2 reference library (*msp1b2RefLib*; section 2.3.4, appendix 7.4) could be used to determine the allelic type of the sequence from which the short reads are derived, as the majority of reads would map to a reference sequence of the same allelic type. This was tested using synthetic reads generated *in silico* from LRD sequences and aligned to the *msp1b2RefLib* using BWA-MEM (Li, 2013). The allelic type of the reference sequence to which the reads were mapped was used to call the correct allelic type for all (964) sequences. Furthermore, there was no significant difference found in the distribution of coverage between the major different allelic types (figure 2.9). This demonstrates that this approach can be used to give unbiased allele calls from short read data for *msp1* block 2. However, this approach cannot be used to determine the presence of MR-type recombinant alleles as these sequences give reads that map to both MAD20- and RO-33-like reference sequences. This problem cannot be overcome by inclusion of MR sequences in the library as reads from both MAD20-like and RO-33-like alleles would map to 5' and 3' ends of these sequences respectively. All 31 MR recombinant sequences present in the LRD contain a conserved sequence (GGTGGTTCAGGTGCTACAGTACCT, from here on known as the MR identifier sequence (MRIS) at the site of recombination between the MAD20- and RO-33-like sequences (figure 2.10). MRIS is unique to MR recombinant alleles and not found in any of the other sequences in the LRD. When synthetic reads were created from LRD MR recombinant sequences and aligned to the *msp1b2RefLib* (section 2.3.4, appendix 7.4), the MRIS was found in at least 9 reads (aligned to either MAD200- or RO-33-like sequences in the *msp1b2RefLib*) for each of the 31 MR recombinant sequences. This suggests that presence of the MRIS can be used to determine the presence of MR alleles from *msp1* block 2 short reads selected by alignment to the *msp1b2RefLib* containing only K1-, MAD20- and RO-33-like sequences.

The MRIS can be used to determine the presence of MR recombinant alleles, but it will not be able to determine whether these are the only *msp1* block 2 alleles present as reads will map to both

MAD20-like and RO-33-like sequences when only MR-recombinant alleles are present or when they are present in a mixed genotype infection with either MAD20-like, RO-33-like or both MAD20-like and RO-33-like alleles. However, if the MR recombinant alleles are not present as part of a mixed genotype infection with MAD20-like or RO-33 like sequences the 3' and 5' ends of these sequences will not be present, as they are lost in the recombination event that forms the MR recombinant alleles (figure 2.10). To test this, synthetic mixed genotype infections were created *in silico* by mixing equal numbers of randomly selected reads generated from LRD MR recombinant sequences and either RO-33 like sequences or MAD20-like sequences or both. Reads were then aligned to the *msh1b2RefLib* (section 2.3.4, appendix 7.4) and, as before, the presence of MR alleles could be detected by searching reads for the MRIS. As predicted, the presence of either MAD20-like or RO-33-like 3' and 5' sequences (figure 2.10) could be used to determine when either or both of these allelic types were present., even when the overall coverage was low (10-fold).

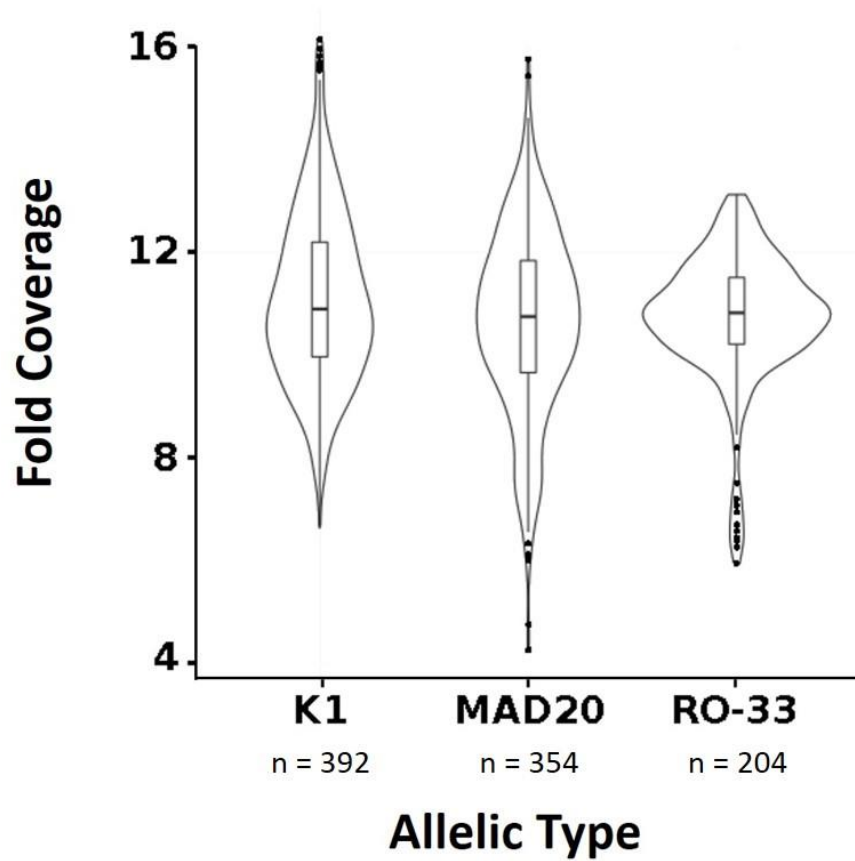


Figure 2.9. Distribution of coverage by allelic type after alignment of synthetic reads to *msp1b2RefLib*. 100 bp synthetic reads created from 964 long read sequences were aligned to the *msp1b2RefLib* of 15 *msp1* block 2 sequences (see section 2.3.4, appendix 7.4) using BWA-MEM (Li, 2013). Coverage was calculated for each alignment and is shown for each major allelic type.

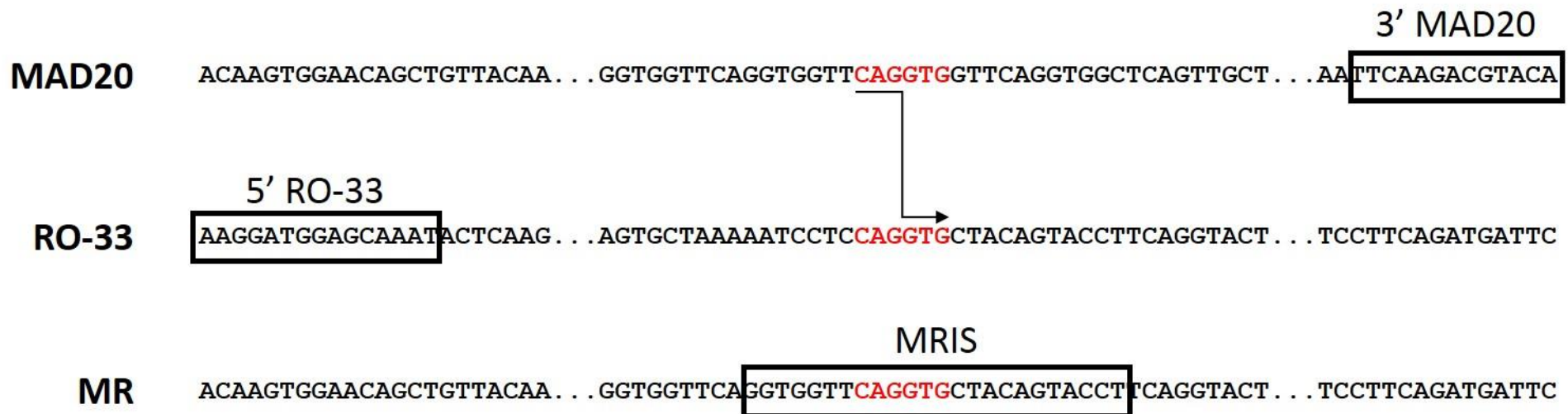


Figure 2.10 Recombination of MAD20- and RO-33-like alleles during formation of MR recombinant allele creates unique sequence. Alignment of a MAD20-like sequence (Accession number AF62449) and the RO-33 sequence (Certa et al., 1987) with the hypothetical recombination site highlighted (red) that produces the MR recombinant allele (Accession number AB502487) containing the conserved, unique MR identifying sequence (MRIS). The MR recombinant allele does not contain the conserved 5' RO-33 sequence (5' RO-33) or the conserved 3' MAD20 sequence (3' MAD20) which can therefore be used to determine the presence of RO-33-like and MAD20-like alleles in combination with MR recombinant alleles in a mixed genotype infection.

2.3.6 Global distribution of *msp1* block 2 alleles as determined from short read data is similar to historical long read data

Aligning short read sequence data from the Pf3k dataset to the *msp1b2RefLib* (section 2.3.4, appendix 7.4) resulted in the alignment of 9.39×10^5 reads from the 2400 isolates collected from sites in West Africa and the Democratic Republic of the Congo (1119 isolates, from here on referred to collectively as West Africa), East Africa (271 isolates) and South East Asia and Bangladesh (1010 isolates, from here on referred to collectively as Asia). In order to allow for comparison between different allelic types with different sequence lengths, coverage, scaled for differences in lengths of reference sequences (appendix 7.4), was calculated for each sample, giving a global mean coverage of 19.3 fold. Coverage of the *msp1b2RefLib* is approximately five times lower than the genome-wide mean coverage of 84.5 fold because reads from one allelic type are mapped across five different sequences in the *msp1b2RefLib*, whereas genome-wide coverage is derived from mapping to a single reference sequence.

One hundred and thirty five samples had been culture adapted prior to sequencing, resulting in a higher coverage (mean coverage 84.1 fold, $p < 0.001$, Wilcoxon signed rank test) which is also seen in the mean coverage genome wide (mean coverage for culture adapted samples is 129 fold compared to 84.5 fold for isolates directly sequenced from patient samples, $p < 0.001$, Wilcoxon signed rank). Comparing the coverage of only directly sequenced isolates shows a significantly higher coverage for African samples (mean coverage 25.7 fold) compared to Asian samples (mean coverage 10.3 fold, $p < 0.001$, Wilcoxon signed rank), which is also seen genome-wide (mean coverage for African samples is 97.5 fold compared to for 65.3 fold for Asian samples). This is due to the lower levels of clinical immunity to malaria meaning that, on average, patients will report symptoms and therefore be diagnosed with malaria, exposing them to the studies contributing samples to the Pf3k, with lower parasitaemias than in Africa.

To determine if there was any bias in the coverage of different allelic types, samples determined to contain only one allelic type (see below) were compared to exclude any bias due to different levels of clones in mixed genotype infections. Due to the variation in coverage between African and Asian samples, comparisons were made between different allelic types within one continent. Comparing between all three major allelic types (K1-like, MAD20-like and RO-33-like) showed no difference in coverage ($p > 0.48$, appendix 7.5).

After determining the presence of MR alleles (see above section 2.3.5) at least one allelic type was found to be present in 2385 (99.4%) of the 2400 clinical isolates with a total 3815 allelic types detected due to mixed genotype infections (table 2.4), of which 1455 (38%) were K1-like, 1384 (36%) were MAD20-like, 860 (23%) were RO-33-like and 116 (3.0%) were MR recombinants (table 2.2). These global frequencies are comparable to those for historic data using PCR-based genotyping (table 2.1). As was found by previous studies, allele frequencies were skewed in favour of K1-like alleles in African populations and in favour of MAD20-like alleles in Asian populations (figure 2.11, $p < 0.001$, χ^2 test) but, with the exception of MR recombinant alleles (see below), allele frequencies did not vary significantly between East and West Africa ($p = 0.23$, χ^2 test). Comparison between typing by alignment of short read data in this study and historic PCR data shows that the allele frequencies are similar, however, the method described here gives a greater skew toward K1-like alleles in West Africa and a greater skew toward MAD20-like alleles in Asia (table 2.3). Allele frequencies do not differ significantly between the data presented here and historic PCR typing studies in East Africa ($p = 0.39$, χ^2 test). There was variation in allele frequencies between study sites within each region, which was starker in Asia than in Africa, likely due to the higher degree of population structure in Asia (figure 2.12).

Region	K1-like	MAD20-like	RO-33-like	MR recombinant
West Africa	897 (0.46)	474 (0.24)	479 (0.24)	109 (0.056)
East Africa	222 (0.44)	144 (0.29)	131 (0.26)	6 (0.012)
Asia	336 (0.25)	766 (0.57)	250 (0.18)	1 (0.00074)
All	1455 (0.38)	1384 (0.36)	860 (0.23)	116 (0.030)

Table 2.2 Allele frequencies determined by alignment to a library of reference sequences. Short read data for 2400 samples from the Pf3k project was aligned to a library of reference sequences (section 2.3.4, appendix 7.4). The presence of reads mapping to a library sequence of one of the three major *msp1* block 2 allelic types (K1-like, MAD20-like and RO-33-like) was used to determine the presence of that allelic type. Presence of MR recombinant alleles were determined by the presence of the MRIS (section 2.3.5, figure 2.10). In the cases where MR recombinant alleles were present, additional presence of MAD20-like and RO-33-like alleles was determined by searching aligned reads for the presence of conserved 3' and 5' sequences (section 2.3.5, figure 2.10). Total counts for each allelic type are shown with fraction of all alleles detected in parentheses.

MR recombinant alleles were detected, by searching reads aligned to the *msp1b2*RefLib (section 2.3.4, appendix 7.4) for the MRIS (figure 2.10), in a total of 116 (4.8%) isolates. In 51 of these fewer than 10 reads containing the MRIS were identified and these alignments were checked by eye to confirm the presence of reads containing MR sequence. As described above, the additional presence of MAD20-like and RO-33-like sequences was determined by the presence of 3' MAD20 and 5' RO-33 conserved sequences. As only a handful of studies have used PCR genotyping to determine MR recombinant allele frequencies, formal comparison was not performed. However, the range of MR recombinant allele frequencies in African sites (2.1-10%) is in agreement with previous studies (figure 2.12; Noranate et al., 2009, Takala et al., 2006, Apinjoh et al., 2015). MR recombinant alleles had higher frequency in West Africa (5.6%) than in East Africa (1.2%), possibly due to the fact that all East African samples in the existing Pf3k dataset were collected in just two sites just over 100km apart. Only one MR recombinant allele was detected in Asia. MR recombinant alleles have previously been reported in Asia (Takala et al., 2006), but there are no large-scale studies that determine their frequency.

Overall, 46% of infections contained two or more allelic types. Increased transmission intensity would predict a higher percentage of mixed genotype infections in Africa compared to Asia, and this was found to be the case with 56% of infections containing one or more allelic types in Africa compared to 31% in Asia ($p < 0.001$, χ^2 test, figure 2.11). This difference is even clearer when considering infection with three or more allelic types which is seen in 19% of infections in Africa compared to just 2% in Asia ($p < 0.001$, χ^2 test); all 32 infections containing all four allelic types occurred in Africa.

Region	Alignment short read data			PCR genotyping			p-value (χ^2 test)
	K1	MAD20	RO-33	K1	MAD20	RO-33	
West Africa	897(46)	474(24)	479(25)	1832(42)	1241(29)	1248(29)	<0.001
East Africa	222(44)	144(29)	131(26)	2127(42)	1453(29)	1470(29)	0.39
Asia	336(24)	766(57)	250(19)	728(35)	1008(48)	367(17)	<0.001

Table 2.3 Comparison of *m*sp1 block 2 genotyping by alignment of short read data and published type-specific PCR. Short read data from the Pf3k project was aligned to a library of *m*sp1 block 2 sequences. Reads mapping to an allelic type were taken to indicate the presence of that genotype in the sample. When MR recombinant alleles were detected, presence of the 3' end of the MAD20 sequence and 5' end of the RO-33 sequence was used to infer presence of the MAD20 or RO-33 alleles respectively. Raw counts for each major allelic type are shown with a percentage of total in parentheses. These data are compared to published PCR-based genotyping from the same region (see appendix 7.1 for a list of all studies). p-values determined by χ^2 test are shown for comparison of short read alignment and PCR genotyping by region.

Number allelic types	Africa	Asia
1	609 (44)	692 (69)
>1	781 (56)	318 (31)
>2	274 (20)	25 (2.5)
4	32 (2.3)	0 (0.0)
Samples	1390 (100)	1010 (100)

Table 2.4 Comparison of mixed genotype infections between Africa and Asia based on *msp1* block 2 genotype. Alignment of short read data from the Pf3k project to a library of K1-like, MAD20-like and RO-33-like *msp1* block 2 was used to determine the presence of these three alleles. Aligned reads were then searched to determine the presence of the unique sequence present in the MR recombinant allele. For samples containing the MR recombinant sequence, the presence of the MAD20 3' sequence and RO-33 5' sequence was used to infer whether this sample contained only MR-like alleles or MR mixed with MAD20-like and RO-33 like. The number of samples containing just one, two or more, three or more or all four allelic types is shown for Africa and Asia with percentage of all 2400 samples shown in parentheses. Mixed genotype infections had higher frequency in Africa than in Asia ($p < 0.001$, χ^2 test).

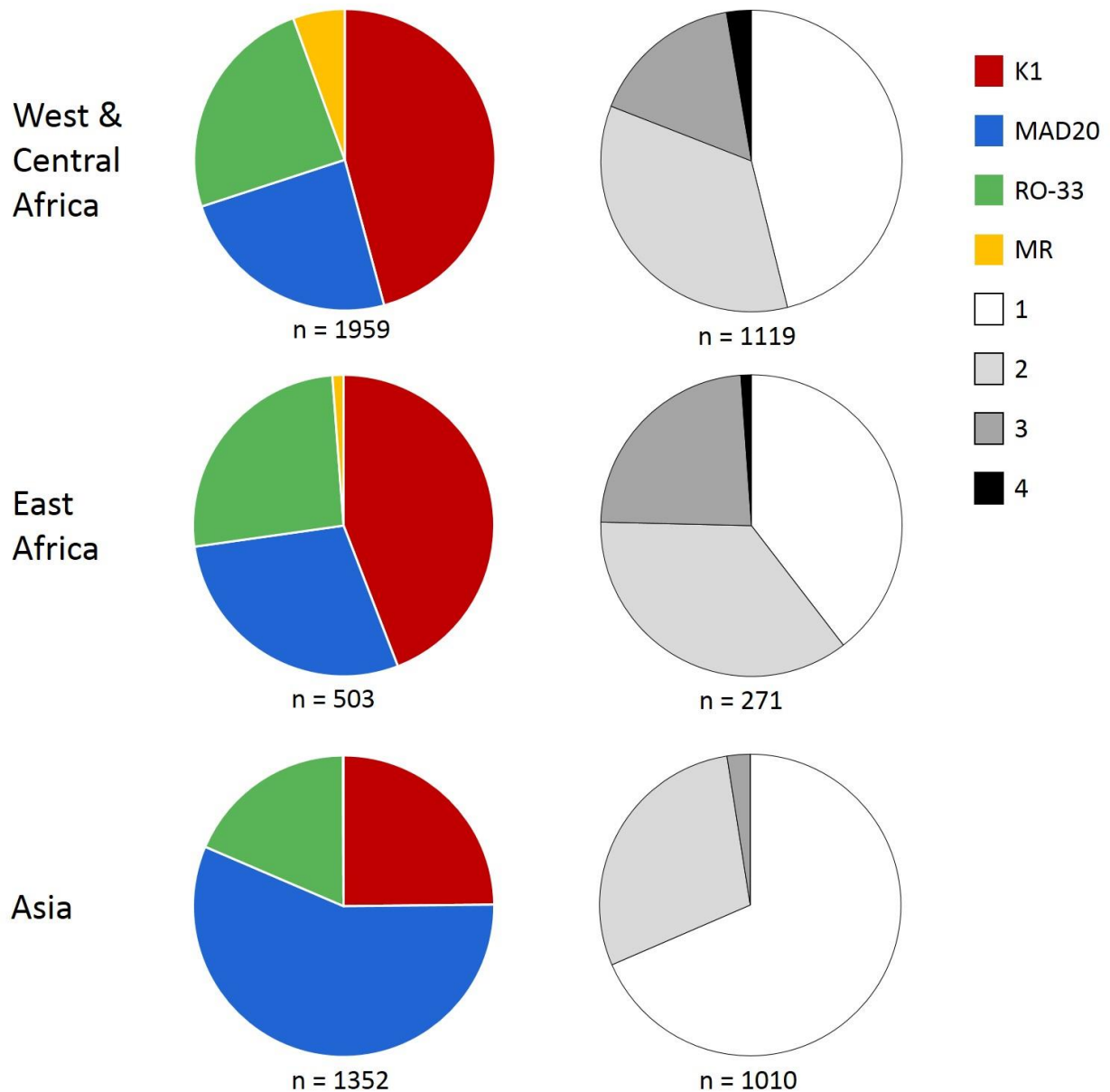


Figure 2.11 Frequencies of alleles and mixed genotype infections vary by region. Alignment of short read data from 2400 Pf3k samples to the *msp1b2RefLib* (section 2.3.4, appendix 7.4) was used to determine the presence of K1- (red), MAD20- (blue) and RO-330-like (green) *msp1* block 2 alleles in each sample. Presence of MR recombinant alleles (yellow) was determined by searching aligned reads for MRIS (section 2.3.5, figure 2.10). The number of different allelic types present in each infection was also determined as either one (white), two (light grey), three (dark grey) or all four (black) depending on the number of *msp1* block 2 alleles detected. In the samples where MR recombinant alleles were detected, the presence of additional MAD20-like and RO-33-like alleles was confirmed by searching aligned reads for conserved 3' and 5' sequences, respectively (section 2.3.5, figure 2.10). The total number of genotypes detected (for pie charts showing allele frequencies) or the total number of isolates (for pie charts showing frequencies of mixed genotype infections) is shown below each pie chart.

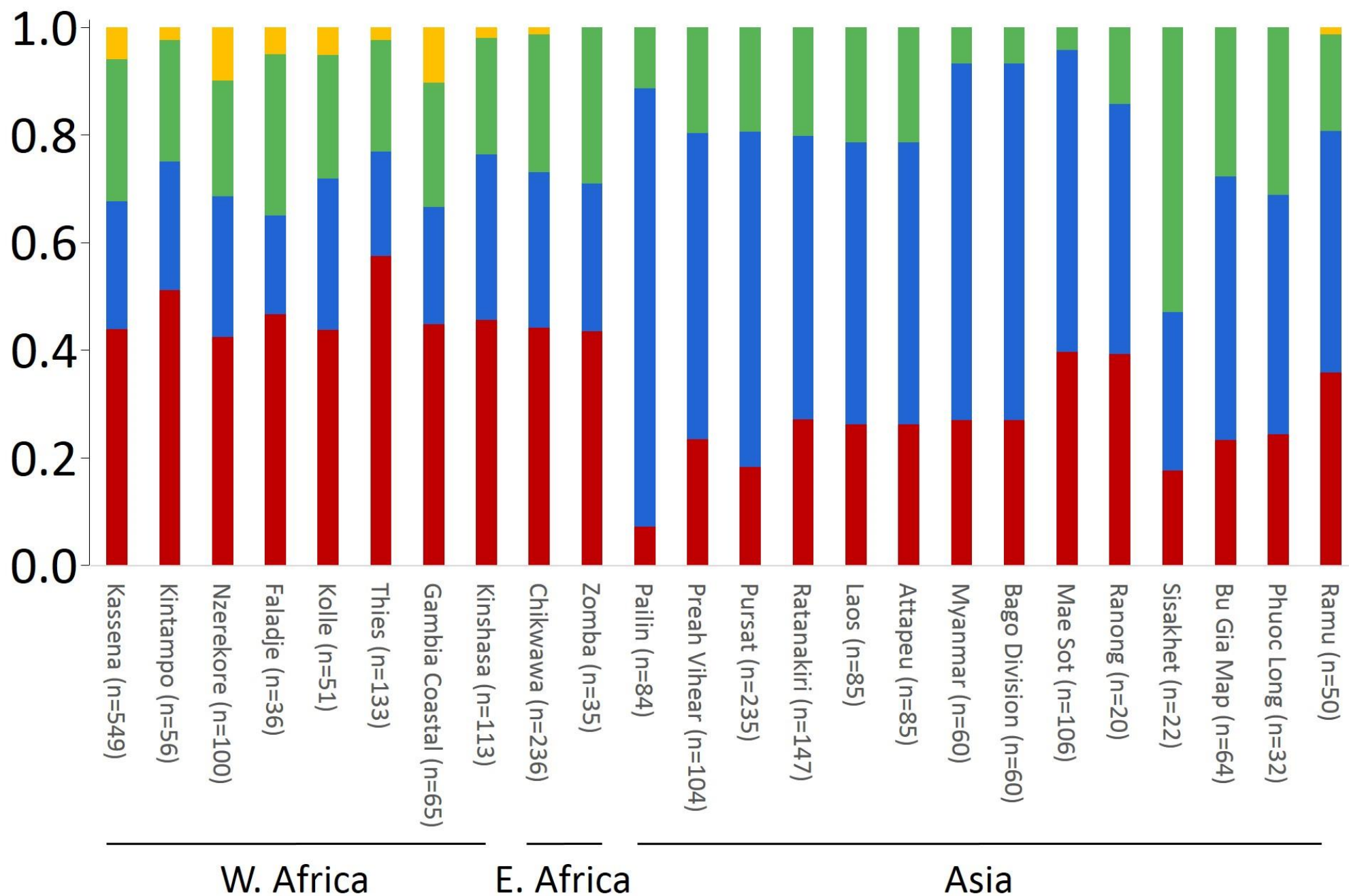


Figure 2.12 Allele frequencies across different study sites. Short read sequence data from the 2400 samples in the Pf3k dataset was used to detect the presence of *msp1* block 2 allelic types by alignment and detection of identifying sequences (sections 2.3.4 and 2.3.5). The frequencies of K1-like (red), MAD20-like (blue), RO-33-like (green) and MR recombinant (yellow) are shown for each site where over 20 samples were sequenced. The sites are grouped by region and the number of samples collected at each site is shown in parenthesis. Map showing the location of all sites contributing samples to the Pf3k project can be found in appendix 7.6.

2.3.7 *De novo* assembly of library-aligned reads increases yield of sequences

Due to the fact that the optimal *k-mer* value was found to be 81 bp (section 2.3.2), only Pf3k samples sequenced with a read length of over 90 bp were used for *de novo* assembly. Reads from these samples that mapped to the *msp1* locus (including 2 kb upstream of the gene) of the 3D7 reference genome were extracted along with their mate pairs (even if the mate pair was unmapped). These reads were then assembled with Velvet using with an optimised *k-mer* value of 81 (section 2.3.2). Assembled contigs were scanned for the presence of a complete block 2 sequence, which was found for 791 (41.9%) out of 1886 samples with read length greater than 90 bp. For the vast majority of these samples only one sequence could be assembled, with just 9 samples giving two sequences. In order to increase the yield of complete block 2 sequences, reads aligned to the *msp1b2RefLib* (section 2.3.4) were used for *de novo* assembly. Using reads mapping to the library dramatically increased number of samples for which sequence was obtained to 1362 (77.2% of all samples with read length > 90 bp); 150 samples gave over two sequences, with 8 samples giving three sequences and one sample giving four. Both sets of *de novo* assembled sequences were collated, removing identical sequences produced by the two approaches, to give a combined 1522 sequences, known from here on as the short read assembled (SRA) dataset (additional data file “Pf3k_short_read_assembled_translated_sequences.csv”). The SRA contains 529 (35.3%), 631 (40.8%), 230 (21.7%) and 32 (2.10%) sequences of the K1-like, MAD20-like, RO-33-like and MR recombinant allelic type respectively with allele frequencies differing between Asia and Africa as expected given the established regional differences in allele frequencies (table 2.1).

For all samples where sequences could be assembled, the allelic type(s) of the assembled sequence matched the allelic type(s) called by alignment to the *msh1b2RefLib* (section 2.3.4) demonstrating the validity and compatibility of both methods. Alignment data cannot determine the presence of two different alleles of the same allelic type as reads from both alleles align to the same group of sequences in the *msh1b2RefLib*. *De novo* assembly produced two or more distinct alleles of the same allelic type in just 11 (0.83%) samples for which sequence was obtained. In all other cases where *de novo* assembly indicated a mixed genotype infection, this was also called by alignment of reads to the *msh1b2RefLib*. As would be predicted from analysis of mixed genotype infections using the alignment of short reads to the *msh1b2RefLib* (table 2. 4), there were approximately three times as many mixed genotype infections detected by *de novo* assembly in Africa (17% of total samples for which sequence was obtained) than Asia (5.5% of total samples for which sequence was obtained).

2.3.8 Agreement between *de novo* assembly and alignment to the *msh1b2RefLib*

In order to explore any bias in the *de novo* assembly of *msh1* block 2 sequences, reads from the Pf3k project were assembled (see above, section 2.3.7). The allelic types of assembled sequences were determined and were compared to the allele calls made by alignment (section 2.3.5). Only samples for which all reads aligned to one allelic type were included in order to avoid a bias against samples from regions with a higher endemicity where both mixed genotype infections and K1-like alleles are at a higher frequency (figure 2.11). No significant difference was found in the allele frequencies between the two procedures ($p = 0.17$, χ^2 test, table 2.6). Cross-checking between the two methods for typing showed that *de novo* assembly never produced a sequence of an allelic type that had not been detected in the given sample by alignment to the *msh1b2RefLib*.

Region	K1-like	MAD20-like	RO-33-like	MR-like
West Africa	269 (0.48)	124 (0.22)	134 (0.24)	28 (0.050)
East Africa	98 (0.42)	61 (0.26)	70 (0.30)	3 (0.013)
Asia	172 (0.23)	436 (0.59)	126 (0.17)	1 (0.0014)
Global	539 (0.35)	621 (0.41)	330 (0.22)	32 (0.021)

Table 2.5 Allelic frequencies of de novo assembled sequences. Reads mapping to the *msp1* locus (including 2kb upstream) and their mate pairs were extracted from samples in the Pf3k dataset which had been sequenced using reads > 90 bp in length. *Msp1* block 2 reads were also extracted by alignment to a library of sequences (section 2.3.4, appendix 7.4). Both sets of reads were assembled independently using Velvet (Zerbino and Birney, 2008) with a *k-mer* length of 81. The two sets of assembled sequences were combined (removing identical sequences) to give a total of 1532 sequences. The number of sequences identified as belonging to each allelic type is shown, with the fraction of all assembled sequences in parenthesis, for the three regions covered by the Pf3k project (West Africa (West Africa and DRC), East Africa and Asia (South East Asia and Bangladesh)) and globally.

	<i>De novo</i> assembly Count (% total)	Alignment Count (% total)
K1-like	251(34.1)	367 (38.6)
MAD20-like	376 (50.7)	457 (48.1)
RO-33-like	108 (14.7)	127 (13.4)
Total	735 (100)	951 (100)

Table 2.6 Comparison of frequencies of allelic types called by alignment and de novo. Reads from Pf3k samples that were homogenous with respect to *msp1* block 2, as determined by alignment to a library of sequences, and had a length of over 90 bp were assembled with Velvet. Allele calls were made based on the complete, assembled sequence. Counts of allelic type are shown with percentage of total called in parentheses. These data are compared to the allele calls made by alignment to a library of *msp1* block 2 sequences for samples that meet the requirements specified above (homogenous at *msp1* block 2 and read length over 90 bp.) There is no significant difference in the distribution of allele frequencies between the two methods for calling allelic type ($p = 0.17$, χ^2 test).

2.3.9 *Msp1* block 2 repeat lengths and structures determined by *de novo* assembly vary between Africa and Asia

With the exception of RO-33-like sequences, the variation between different *msp1* block 2 alleles is almost entirely due to variation in the repeat sequences (Noranate et al., 2009). These changes in nucleotide sequence invariably lead to changes in amino acid sequence. Due to this fact, and the interest in these sequences as immune epitopes, the *de novo* assembled nucleotide sequences present in the SRA dataset were translated into their predicted amino acid sequences to give a dataset of peptide sequences, the translated SRA (tSRA) dataset. Analysis of the amino acid sequences shows that K1-like sequences are the most diverse with 224 distinct peptide sequences (from here on referred to as alleles) compared to 123 different MAD20-like, 9 RO-33 and just 6 MR recombinant alleles. The number of alleles within both the K1-like and MAD20-like allelic types was greater in African populations than in Asian populations (table 2.7). This is despite the fact that MAD20-like alleles are at a far higher frequency in Asia ($p < 0.001$, χ^2 test, table 2.5). *De novo* assembly of the Pf3k data revealed 246 novel alleles that had not been seen by previous studies that deposited *msp1* block 2 sequences in GenBank (additional data file “Pf3k_short_read_assembled_translated_sequences.csv”). The vast majority of these, 166 (67%) were of the K1-like variety with 73 (30%) MAD20-like, five (2.0%) RO-33-like and two (0.81% %) MR recombinant.

Region	All types	K1-like	MAD20-like	RO-33-like	MR-like
West Africa	240 (555)	166 (269)	60 (124)	8 (134)	6 (28)
East Africa	101 (230)	65 (98)	32 (61)	3 (70)	1 (3)
Asia	74 (735)	22 (172)	48 (436)	3 (126)	1 (1)
Global	363 (1522)	225 (539)	123 (621)	9 (330)	6 (32)

Table 2.7 Number of unique peptide variants by region. *Msp1* block 2 short read sequences from the Pf3k data set were assembled using Velvet (Zerbino and Birney, 2008) with *k-mer* length of 81. Contigs containing complete *msp1* block 2 sequences were compiled and translated to the predicted amino acid sequence. Unique peptide sequences (referred to as alleles) were determined by alignment. Total unique alleles are shown with the total number of sequences of that allelic type in parenthesis for each region covered by the Pf3k project (West Africa and DRC (West Africa), East Africa, and South East Asia and Bangladesh (Asia)).

The tripeptide repeats encoded by the K1-like alleles consist of four major tripeptides: serine-alanine-glutamine (SAQ), serine-glycine-alanine (SGA), serine-glycine-threonine (SGT) and serine-glycine-proline (SGP). The repeat almost always begins with the SAQ tripeptide, which, if present more than once, is always repeated as part of a motif with one of the other three tripeptides (resulting in SAQSGA, SAQSGT or SAQSGP motifs). Likewise, the SGA tripeptide is only present as part of the SAQSGA motif. SGT and SGP tripeptides are commonly encoded at the 3' end of the repeat sequence and can be repeated as part of a motif (for example SGTPSGPSGTSGP) or independently (for example SGTSGTSGT) (Miller et al., 1993). Additionally rare K1-like *msh1* block 2 variants encoding serine-alanine-threonine/proline (SAT/P) tripeptides have been recorded (Juliano et al., 2010) but were not seen in the *de novo* assembled data presented here.

Comparison of K1-like tripeptide repeat lengths showed that African alleles (median length: 12 tripeptides) tended contain a greater number of tripeptides than Asian alleles (median length: 8 tripeptides, $p < 0.001$, Wilcoxon signed rank test); a wider range in the number of tripeptide repeats was seen in Africa (seven to 19 tripeptides) than in Asia (five to 16 tripeptides). This is one factor explaining the increased number of K1-like variants seen in Africa.

Analysis of the tripeptide repeat composition reveals that all but 3 of the 172 (98.2%) Asia alleles do not have any SGA motifs in the tripeptide repeat and the three that do are all identical sequences from the same site in Bangladesh. This is in contrast to Africa where the SGA motif is present in 211 (57.5%) of all K1-like tripeptide repeats. All of these alleles encode a tripeptide repeat starting with a SAQSGA motif (meaning they can be classified within the 3D7-like subtype), which is repeated in just over half (55.9%) of these sequences between one and three times. This is consistent with previous data that has identified a high frequency of alleles containing the SAQSGA motif in Africa (Noranate

et al., 2009) but their absence in South East Asia (Tanabe, 2013). Interestingly, this motif is present in a previously published *msh1* block 2 sequence from North Eastern India (Joshi et al., 2007).

With the exception of six alleles, all the Asian K1-like alleles encode a tripeptide repeat consisting of a single SGA tripeptide followed by between two and eight SGT tripeptides and then either SGPSGPGT (122/164 samples) or SGPSGTSGT (36/164 samples). This repeat structure (which is shared by the K1 isolate and K1-like alleles of the K1-like subtype) is seen in Africa as part of a mix of a much wider range of repeat structures. Amongst African alleles, SGA tripeptides are always part of an SAQSGA motif which, when repeated, is always encoded at the 5' end of the repeat sequence and is not interspersed with SGT or SGP-containing motifs. In general, the 5' end of the K1-like of the tripeptide repeat encodes SAQSGA motifs (if present) and the 3' end of the repeat encodes SGT and SGP tripeptides, but SAQ tripeptides can be found in all but the final two positions of the tripeptide repeat sequence. This heterogeneity in repeat structure, combined with the increased range of tripeptide repeat lengths results in a far greater number of K1-like alleles detected in Africa compared to Asia, even when considering the increased number of African K1-like sequences (table 2.7).

Despite the large number of K1-like alleles present in Africa, the top five most abundant K1-like alleles collectively account for 18.8% (69/367) of K1-like sequences assembled (figure 2.13) and appear closely related. All five alleles encode tripeptide repeats consisting of either one, two, three or four 5' SGASQA motifs followed by one or four SGT tripeptides then two or three SGP tripeptides before the final SGT tripeptide (figure 2.16), suggesting they arose by repeat expansion or retraction from a shared ancestor. One of these alleles is identical to the only SGASQA containing allele in Asia, but all other alleles are unique to Africa and found across all African sites sampled. The many other African alleles are all present at low frequency (< 2% of K1-like sequences). In Asia just 8 alleles account for 86.6% (149/172) of all K1-like sequences. Again, these alleles share a repeat structure, consisting of three to 8 SGT tripeptides followed by one or two SGP tripeptides and then either one

or two SGT tripeptides (figure 2.16), suggesting they originate via repeat expansion or contraction from a single common ancestor. These alleles are found across all Asian sites and five of them are also found in Africa, but at low frequencies (figure 2.13).

There were two African alleles encoding K1-like tripeptide repeats in Africa that did not fit the general pattern described above. One sample from Malawi contained an allele encoding a tripeptide repeat where the first SGA tripeptide had been lost. Another sample from Malawi has a K1-like allele that is missing 60 bp from the 3' non-repeat sequence. There is a rare SNP in the 5' non-repeat sequence which is present in 19/367 African K1-like samples which were assembled. None of these rare variants have been reported previously. The SAT/P tripeptides, originally reported to be encoded in the sequence of the 3D7 *msh-1* gene (Miller et al., 1993) but subsequently shown to be absent from the 3D7 sequence, and also reported in a small number of field isolates from both Africa and Asia (Juliano et al., 2010, Joshi et al., 2007) are absent from all of the 539 K1-like sequences assembled here.

The MAD20-like repeats are known to be comprised of five different tripeptides: serine-lysine-glycine (SKG), serine-glycine-glycine (SGG), serine-valine-alanine (SVA), serine-serine-glycine (SSG) and serine-valine-threonine (SVT) (Miller et al., 1993). The SKG, only occurs as the first tripeptide, only ever occurs once in the repeat and is at a higher frequency in Africa (63.7% of MAD20-like sequences) than in Asia (39.4%, $p < 0.001$, χ^2 test). When SKG is the first tripeptide, it is always followed by either SVA, SVT or SGG, in Asia all three motifs are common, but in Africa SVT is at a higher frequency than in Asia and SGG is rare ($p < 0.001$, χ^2 test). The first position in the repeat sequence can also be occupied by serine-glycine-glycine SGG or serine-valine-alanine SVA, which can both occur multiple times in the rest of the repeat sequence. The serine-serine-glycine SSG tripeptide is almost exclusively found in Asian MAD20-like alleles and only occurs once. The SSG tripeptide is present in 57 (13.1%) of the 436 MAD20-like sequences assembled from Asian samples, the majority (51/57) have this tripeptide as the first tripeptide followed by varying numbers of SGG

and SVA repeats. The other 6 samples all contain the same allele, which has the SGG tripeptide in the middle of the tripeptide repeat; this allele has previously been reported at low frequency in Asia and in one East African isolate (Juliano et al., 2010). Two African alleles analysed here contain a serine-aspartic acid-glycine (SDG) tripeptide, which has not been previously reported, but was present in one allele sequenced by Juliano et al (2010).

MAD20-like tripeptide repeats have a greater range of lengths in Africa (5 to 19 tripeptides) than in Asia (5 to 17 tripeptides) and African alleles tend to encode longer stretches of tripeptide repeats (median = 14 tripeptides) than Asian alleles (median = 11 tripeptides, $p < 0.001$, Wilcoxon signed rank test), resulting in the greater number of alleles seen in Africa. This appears to be the result of expansion of SGG tripeptide repeats which, on average, comprise a greater fraction of MAD20-like repeat sequences in Africa (median 58.3% of tripeptides in repeat) than in Asia (median 50.0%, $p < 0.001$, Wilcoxon signed rank test). The second most common tripeptide, SVA, comprises the a similar fraction of the tripeptide repeat, on average, in both Africa (median 33.3%) and Asia (median 36.4%). However, in Africa this tripeptide is always followed by the SGG tripeptide whereas in Asia many alleles, including three of the five most abundant alleles, contain tandemly repeated SVA tripeptides. Accordingly, the length of the MAD20-like tripeptide repeat shows a stronger correlation to the number of SGG ($r = 0.90$, 95% confidence interval (CI) [0.87,0.92], $p < 0.001$, Pearson's correlation) tripeptides than the number of SVA tripeptides in Africa ($r = 0.53$ 95% CI [0.42,0.63], $p < 0.001$, Pearson's correlation) whereas in Asia, increases in SGG ($r = 0.82$, 95% confidence interval (CI) [0.78,0.85], $p < 0.001$, Pearson's correlation) and SVA ($r = 0.80$, 95% confidence interval (CI) [0.77,0.83], $p < 0.001$, Pearson's correlation) tripeptides contribute equally to the increased length of the repeat.

Determining the frequency of each MAD20-like allele amongst all MAD20-like sequences from each continent reveals a similar pattern to that seen for K1-like alleles; the most common MAD20-like alleles in one continent are significantly less frequent or absent from the other, with the exception

of one allele which is found at comparable frequencies in both (figure 2.13) . Again, two alleles in Asia are present in over 10% of samples for which a MAD20-like sequence could be assembled, whereas in Africa there are no alleles above this frequency but a greater number of alleles at lower frequencies.

Analysis of RO-33-like sequences *de novo* assembled from the Pf3k data showed the presence of 7 single nucleotide polymorphisms (SNPs). One of these SNPs, found in just one Asian sample was synonymous, whilst the remaining 6 SNPs resulted in an amino acid substitution. At one position two variant bases mean that there are three different alleles creating 7 variants of the RO-33 peptide sequences that each contain one amino acid difference from the RO-33 sequence. The exact match to the RO-33 allele sequence accounted for over three quarters (155/204) of all assembled RO-33-like sequences from Africa but was found at very low frequency (3/126 assembled RO-33-like sequences) in Asia (figure 2.14), in agreement with previous studies (Noranate et al., 2009, Tanabe et al., 2013). Just under one fifth (37/204) of African RO-33-like sequences had a SNP at the 3' end of block 2 encoding substitution of a glycine for an aspartic acid residue (G97D), which was absent from Asia (figure 2.14). Another SNP towards the 3' end of block 2 encodes an amino acid change from lysine to asparagine (K90N) and is found at low frequencies in both Africa (4/204) and Asia (1/126). This mutation has been previously reported (Noranate et al., 2009) but analysis of sequences assembled in this study uncovered a novel variant at this position resulting in mutation to a threonine (K90T). Three additional low frequency SNPs were identified in Africa (G91D, S73N and A74D), two of which (S73N and A74D) have not been previously reported (see "Pf3k_short_read_assembled_translated_sequences.csv"). There was one RO-33-like SNP encoding an aspartic acid to glycine substitution (D67G) that was present in two alleles, one of which contains no other SNPs and accounts for over 95% (122/126) of RO-33-like alleles in Asia but is not found in Africa (figure 2.14). This is in agreement with previous studies that find this allele be the dominant RO-33-like allele in South East Asia (Tanabe et al., 2013, Juliano et al., 2010). Interestingly, this SNP is seen in an African allele for the first time, but in combination with another low frequency SNP

(G91D) that is unique to Africa (figure 2.14) (Noranate et al., 2009). The previously reported Q27E mutation (Noranate et al., 2009) was not found here.

RO-33-like alleles do not contain polymorphic repeat sequences, allowing alignment and analysis of mutations for signatures of selection. Such analysis was performed on the seven SNPs present in the 330 RO-33 like sequences and showed no evidence of selection (nucleotide diversity (π) = 5.7×10^{-3} , Tajima's D = -0.64 ($p > 0.1$), Fu and Li's D = -0.84 ($p > 0.1$), Fu and Li's F = -0.92 ($p > 0.1$)). The same result is found with the 6 SNPs in the 231 African sequences ($\pi = 1.1 \times 10^{-3}$, Tajima's D = -1.70 ($p > 0.05$), Fu and Li's D = -0.93 ($p > 0.1$), Fu and Li's F = -1.40 ($p > 0.1$)) and the 2 SNPs in the 126 Asian sequences ($\pi = 0.055 \times 10^{-3}$, Tajima's D = -1.22 ($p > 0.1$), Fu and Li's D = -1.1 ($p > 0.1$), Fu and Li's F = -1.3 ($p > 0.1$)). This finding is in agreement with previous studies of African RO-33 sequences (Noranate et al., 2009).

The 32 MR-recombinant sequences that were assembled contained just 6 different alleles. Five of these appear to have arisen from the same recombination event as they have identical 5' MAD20-like and 3' RO-33-like sequences and only differ as a result of contraction or expansion of the SGG tripeptide repeat (figure 2.15). This one recombination event between a MAD20-like and RO-33-like allele appears to be responsible for all but one of the MR recombinant alleles that has been sequenced previously (Takala et al., 2002, Noranate et al., 2009)³. It is interesting to note that the 3' sequence present in all MR recombinant alleles is identical to an RO-33-like sequence (bearing the G97D mutation) which is the second most abundant in Africa, but absent from Asia; the one Asian MR recombinant sequence also contains this motif, suggesting that this allele arose via recombination in Africa and was subsequently introduced into Asia. One sample, from West Africa, was found to contain a MR-like allele with a notably different MAD20-like sequence at the 5' end, suggesting this allele arose from an independent recombination between a MAD20-like sequence

³ The sequences published by Takala et al (2002) were amplified using a 3' RO-33 primer that changes the sequence of the 3' MR-recombinant sequence from that found here and in the study by Noranate et al (2009) which uses a 3' primer binding outside the block 2 sequence. All 5' sequences published by Takala et al (2002) match with the majority of 5' sequences in Noranate et al (2009) and those assembled in this study.

and the same RO-33-like sequence, or recombination of an MR-recombinant allele derived from the original MAD20-like/RO-33-like recombination and a different MAD20-like sequence. Long read sequence data also detected one MR-like allele with a divergent 5' end, although this does not match the allele found here (Noranate et al., 2009).

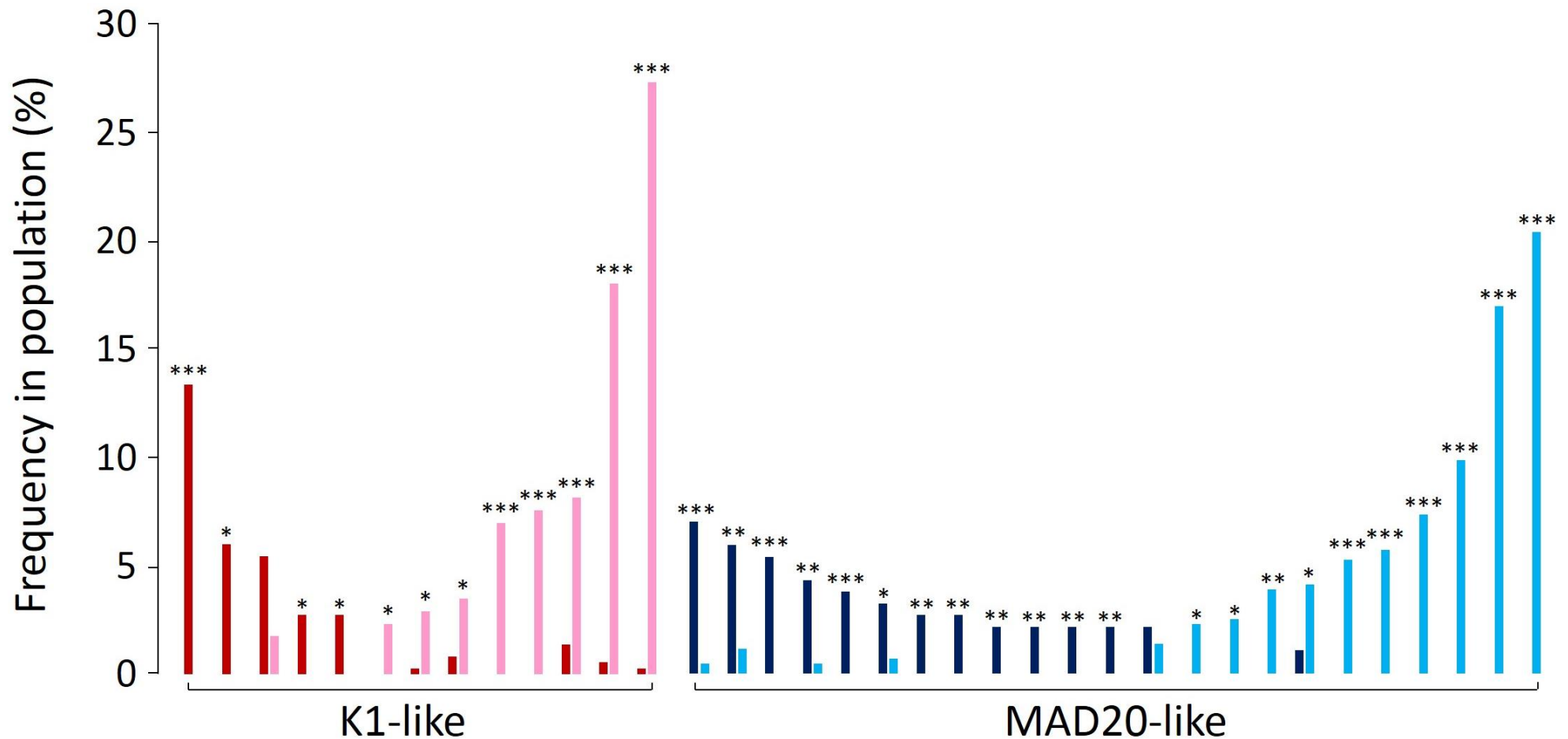


Figure 2.13 Allele frequencies of K1-like and MAD20-like alleles show skews between continents. *Msp1* block 2 allele sequences were determined by *de novo* assembly of short read sequences (section 2.3.7). The frequency as a percentage of all alleles in the population of the same allelic type that had been *de novo* assembled was calculated for both Africa (367 total K1-like sequences; 185 total MAD20-like sequences) and Asia (172 K1-like sequences; 436 MAD20-like sequences). All K1-like sequences (left) that had an allele frequency greater than 2% in either Africa (red) or Asia (pink) are shown; an additional 218 alleles were present at frequencies below 2%. Similarly, all MAD20-like sequences with an allele frequency greater than 12% in either Africa (dark blue) or Asia (light blue) are shown; an additional 103 alleles occur at frequencies below 2%. Asterisks indicate significant differences between the allele frequency in Africa and Asia with p-values < 0.05 (*), < 0.01, (**) or p < 0.001 (***).

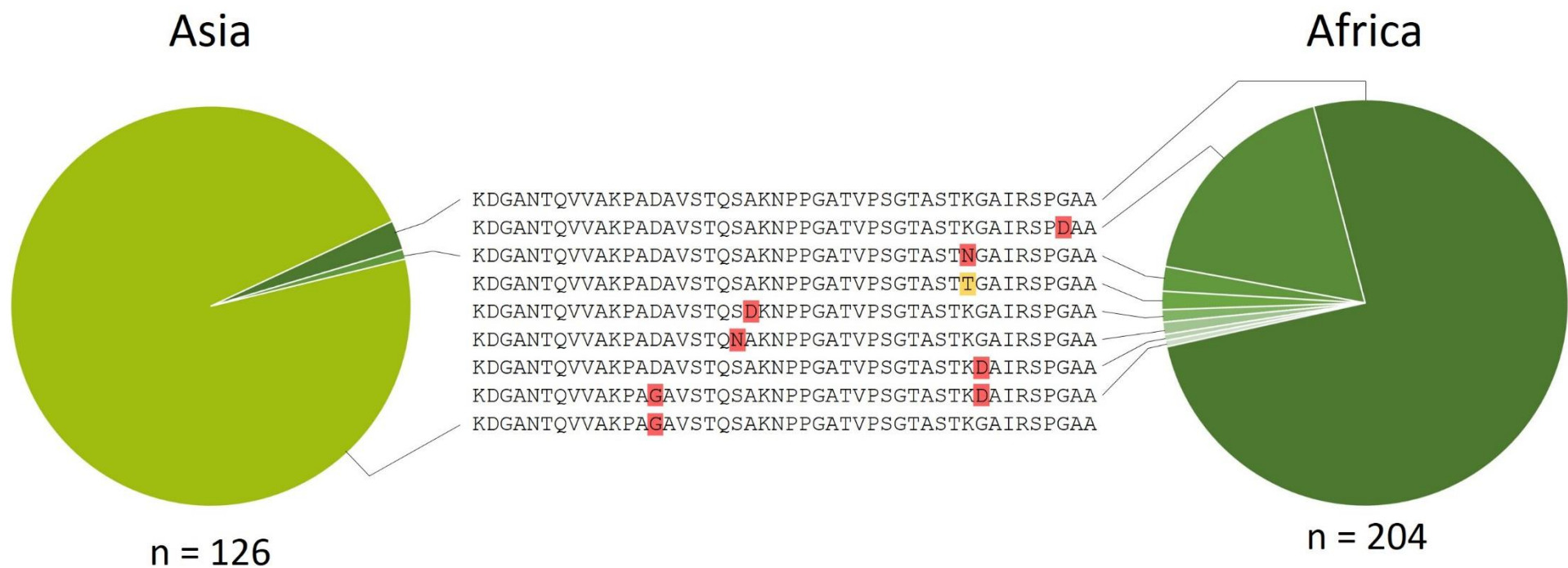


Figure 2.14 RO-33-like alleles and their frequencies. *De novo* assembly of short read sequences mapping to *msp1* block 2 (section 2.3.7) revealed 6 non-synonymous single nucleotide polymorphisms (nsSNPs) in RO-33-like alleles creating 9 different peptide sequences, which are aligned. Changes in peptide sequence are highlighted (yellow and red) and the frequency of each allele is shown in Africa (right) and Asia (left).



Figure 2.15 MR alleles and their frequencies. *Msp1* block 2 sequences were assembled from short reads using Velvet. Nucleotide sequences encoding MR recombinant alleles were translated into the predicted amino acid sequence. Six different alleles, shown here, were found in the 32 samples bearing an MR recombinant allele. The difference between the alleles is due to expansion and contraction of the serine-glycine-glycine (SGG) tripeptide (highlighted in purple) or, in one case, suspected recombination with a MAD20-like allele with a divergent sequence (red). The number of times each allele was found in Africa is shown (right hand side of figure). There was only one sample with an MR recombinant allele in Asia, which was identical to one of the African alleles (indicated with an asterisk).

Sequence

% K1

A

EEITTKGASAQSGASAQSGASAQSGASQ-----SGT-----SGPSGP---SGTSPSSRSNTLPRSNTSSGASPPADA	5.99
EEITTKGASAQSGASAQSGASAQSGASAQSGT-----SGPSGP---SGTSPSSRSNTLPRSNTSSGASPPADA	5.45
EEITTKGASAQSGASAQSGASAQ-----SGT-----SGPSGP---SGTSPSSRSNTLPRSNTSSGASPPADA	2.72
EEITTKGASAQSGASAQ-----SGT-----SGPSGPSPSGTSPSSRSNTLPRSNTSSGASPPADA	2.72
EEITTKGASAQSGASAQSGTSAQSGTSAQ-----SGTSGTSGTSGTSGPSGP---SGTSPSSRSNTLPRSNTSSGASPPADA	1.91

B

EEITTKGASAQSGTSGTSGT-----SGPSGP---SGTSPSSRSNTLPRSNTSSGASPPADA	27.3
EEITTKGASAQSGTSGTSGTSGTSGT-----SGPSGP---SGTSPSSRSNTLPRSNTSSGASPPADA	18.0
EEITTKGASAQSGTSGTSGTSGT-----SGP---SGTSGTSPSSRSNTLPRSNTSSGASPPADA	13.4
EEITTKGASAQSGTSGTSGTSGTSGTSGTSGTSGPSGP-----SGTSPSSRSNTLPRSNTSSGASPPADA	8.14
EEITTKGASAQSGTSGTSGTSGTSGTSGT-----SGPSGP---SGTSPSSRSNTLPRSNTSSGASPPADA	7.56
EEITTKGASAQSGTSGTSGT-----SGP---SGTSGTSPSSRSNTLPRSNTSSGASPPADA	6.98
EEITTKGASAQSGTSGTSGTSGT-----SGPSGP---SGTSPSSRSNTLPRSNTSSGASPPADA	2.90
EEITTKGASAQSGTSGTSGTSGTSGTSGTSGTSGTSGTSGPSGP---SGTSPSSRSNTLPRSNTSSGASPPADA	2.33

Figure 2.16 Alignment of K1-like sequences suggests shared origin for most abundant alleles on each continent. Short (90-100 bp) reads mapping to *msp1* block 2 were extracted and assembled using Velvet. The frequency of each unique alleles was determined by comparison with all assembled sequences. K1-like alleles with a frequency of over 2% in the population of all K1-like alleles in either Africa (A, n= 367) or Asia (B, n=172) were aligned. The proportion of all sequences represented by of each allele in either Africa (A) or Asia (B) is shown to the right of the sequence as a percentage.

2.4 Discussion

In this chapter two approaches were developed to allow the extraction of repetitive sequence data from short sequence reads. The first approach utilised the *de novo* assembly algorithm Velvet and the second approach used a *msp1b2RefLib* representing the range of MSP-1 block 2 sequences to align short reads. The combination of these two approaches allowed for typing of all sequences present in the Pf3k dataset and the assembly of over 700 novel MSP-1 block 2 sequences.

It was shown that, by increasing the *k-mer* length to 81, the DBG-based assembler Velvet (Zerbino and Birney, 2008) can assemble short reads from even long *msp1* block 2 sequences in agreement with both the theory behind de Bruijn graphs and work using ideal data that shows an increased *k-mer* length results in increased assembly (albeit with increased computational costs) (Li et al., 2012). However, the longer the repeat region of original *msp1* block 2 sequence, the less likely that algorithms for *de novo* assembly will be able to resolve the sequence (figure 2.5), as would be expected (Li et al., 2012, Leggett et al., 2013). It is important to note however, that, due to the imperfect nature of the repeat sequences of *msp1* block 2 (Miller et al., 1993), whilst the repeat region may expand beyond the *k-mer* length the presence of some unique sequence within the repeat will mean that the repeat lengths “seen” by the assembler are a fraction of this length (Li et al., 2012). The probability of sequence assembly increases with depth of coverage (figure 2.5), which is expected as the *k-mer* length is close to the length of the repeat sequence meaning that increasing the number of reads (i.e. the coverage) will increase the probability of reads containing *k-mers* that span the repeat sequence and allow resolution of the de Bruijn graph. This is in agreement with modelling of assembly algorithms using perfect data that shows increased assembly with increase of coverage depth (Li et al., 2012). With a coverage depth of over 80-fold assembly of an *msp1* block 2 sequence containing a long perfect repeat was always achieved using ideal data. However, when using real data the coverage depth will need to be higher due to the presence of sequencing errors and multi-clone infections which essentially dilute the number of reads that can be used for

assembly. One approach to reduce bias in *de novo* assembly of repeat regions would be to exclude samples with coverage below a certain threshold. However, this is impractical as coverage varies across the genome, so the coverage at any given locus may differ dramatically from the genome wide mean. When considering just the locus of interest, given that the original sequence length and number of alleles present is unknown, a true estimation of the coverage is impossible to obtain.

By building a library of *msp1* block 2 sequences from published long read sequence data (sequences used are listed in appendix 7.4) it was possible to align short (30-100 bp) reads and determine the allelic types present 99.4% of isolates. The fact that the allele frequencies determined by alignment to a sequence library do not differ significantly from those determined by PCR genotyping in East Africa (which has the greatest number of historical data) and that shifts in allele frequencies from PCR data are seen in favour of both K1-like and MAD20-like alleles suggest there is no systematic bias in the method developed here (table 2.3). The changes in allele frequencies are unlikely to be due to actual shifts as these have been shown to be stable over time (Tanabe et al., 2007a, Noranate et al., 2009, Silva et al., 2000). It is probable that decreased sensitivity of alignment of short read sequence data to detect alleles at low frequencies in mixed genotype infections has resulted in fewer of the less frequent alleles being detected in both West Africa, where there was a small but significant increase in the frequency of K1-like alleles in comparison to historical PCR data, and Asia, where there was a greater increase in MAD20-like alleles. The larger shift in allele frequencies relative to PCR data in Asia is likely due to a much higher degree of heterogeneity between study sites (figure 2.12), caused by increased population isolation due to ecological barriers to transmission in this region (Pumpaibool et al., 2009, Anderson et al., 2000) compared to Africa (Duffy et al., 2017, Mobegi et al., 2014). Mixed genotype infections are common in *P. falciparum*, especially in areas of high endemicity (Anderson et al., 2000). Accordingly, infections containing multiple allelic types were at a much higher frequency in African populations compared to Asian populations (table 2.4; figure 2.11). The replication of findings made using PCR-based genotyping methods clearly demonstrates the validity of the approach developed here to call *msp1* block 2

genotypes using short read sequence data from whole genome sequencing studies and encourages the use of this approach both for future analysis of *msp1* block 2 genotypes and other polymorphic repeat sequences in the *P. falciparum* genome. Unlike PCR amplification, this approach does not yield the length of the repeat sequence, preventing estimations of the multiplicity of infection (MOI). However, estimates of MOI can be made from whole genome data, which will have a higher accuracy than those based on a single locus (Assefa et al., 2014, Murray et al., 2016).

Determination of the presence of rare MR recombinant alleles (Takala et al., 2006) is important if this approach is to be used for surveying changes in allele frequencies resulting from vaccination with MSP-1 block 2 antigens as increased immune pressure on alleles present in the vaccine may result in selection for other alleles. The presence of unique and conserved sequence at the site of the recombination between MAD20-like and RO-33-like sequences (Takala et al., 2002) allows for the detection of these alleles from reads aligned to a *msp1b2RefLib* (section 2.3.5). The use of specific sequences to distinguish the presence of recombinant and allelic types from mixed genotype infections further commends the use of libraries of reference sequences for genotyping complex polymorphic repeat sequences using short read data.

The use of *de novo* assembly not only allows determination of allelic type length but also gives the structure of the repeat sequence. Targeted *de novo* assembly was first attempted using *msp1* block 2 sequence reads mapped to the *msp1* locus of the 3D7 reference genome. It was predicted that polymorphic reads that did not align to the reference genome could still be captured as their mates would be mapped to the conserved sequences flanking block2. However, a far better outcome of targeted *de novo* assembly is achieved if reads are first aligned to a library of polymorphic sequences prior to assembly due to the increased read depth that will aid the assembly of complex regions (Li et al., 2012) figure 2.5). Comparison of the ratios of allelic types in single-genotype infections determined by *de novo* assembly and alignment of short reads shows that *de novo* algorithm used does not result in a bias towards any one allelic type (table 2.6) thus encouraging the use of this

approach for the study of repeat sequences. However, it should be noted that *de novo* assembly will fail when the repeat length is longer than the read length (Li et al., 2012) figure 2.2), and therefore this approach is limited by the length of the repeat and sequencing read.

The targeted *de novo* assembly of *msp1* block 2 reads described here has generated by far the largest single dataset of *msp1* block 2 sequences, yielding a greater number of sequences than all studies so far have deposited in GenBank. Analysis of sequences obtained by targeted *de novo* assembly resulted in the identification of 224 K1-like and 123 MAD20-like unique alleles, the majority of which were present only in Africa (table 2.7), in agreement with previous studies that show decreased diversity at this locus in Asia (Tanabe et al., 2007b). The number of unique alleles identified here is an order of magnitude greater than that determined by fragment size polymorphism in previous studies (Branch et al., 2001, Takala et al., 2006). This is due to polymorphism in both the repeat length and the sequence of the repeat; variation in length of block 2 sequence is far lower than the variation in the sequence. This is relevant to consideration of vaccine design, as repeat expansion will not necessarily create novel epitopes, whereas variation in the repeat sequence will always change the epitopes presented to the immune system.

Analysis of the wealth of sequence data generated by *de novo* assembly of *msp1* block 2 sequences allowed the analysis of variation in the encoded tripeptide repeats of K1-like, MAD20-like and MR recombinant alleles. This is the first dataset encompassing large numbers of clinical isolates from both Africa and Asia and thus allows direct comparison of sequences between the two major zones of malaria transmission. The diversity of all allelic types was far greater in Africa compared to Asia (table 2.7). However, in the case of MAD20-like alleles there was a greater diversity of repeat structures in Asia and the increased numbers of distinct alleles in Africa was driven by the expansion of a single tripeptide repeat (section 2.3.9). This is in contrast to K1-like alleles which in Asia almost exclusively encode a single tripeptide repeat structure but in Africa encompass a wide range of repeat structures and lengths, leading to the astounding diversity recorded here and in other studies

(Noranate et al., 2009). The stark differences in the tripeptide repeat structures between the continents have implications for vaccine design as it may be the case that a formulation containing a greater number of K1-like repeat structures would be optimal for Africa, whereas one with a greater number of MAD20-like repeat structures would perform better in Asia (see chapter 3).

The efficiency of parallel sequencing enables generation of sequence data for a large number of parasite isolates. Targeted *de novo* approaches can harness this data to provide information on highly polymorphic regions and reveal rare variants, as has been shown here for all *msp1* block 2 allelic types. The error correcting algorithms present in DBG based assemblers remove erroneous bases present in a small number of reads due to sequencing errors. This is an advantage over PCR-based sequencing where errors are harder to detect. *De novo* assembly is, however, limited by its dependency on high read depth (figure 2.5). This sets a threshold for the quality of sequencing, but also means that sequences present at low levels in mixed genotype infections will not be assembled as is reflected in the much lower frequency of mixed genotype infections detected by *de novo* assembly (section 2.3.8) as opposed to alignment (section 2.3.5) of short reads despite the potential for *de novo* assembly to detect mixed genotype infections containing different *msp1* block 2 alleles of the same allelic type. The use of coloured de Bruijn graphs has been proposed to enhance the assembly of different alleles from mixed genotype infections (Iqbal et al., 2012). It would be of interest to see if this approach can aid with assembly of alleles present at low frequency although this algorithm has not yet been released publically. NGSreper is a new algorithm that has been developed to assemble repeat sequences from short read data by using identical reads that are present at a frequency above the average read depth to identify and assemble repeat sequences (Lian et al., 2016). Whilst this algorithm has been developed to assemble long repeat sequences, it would be of interest to test this with *msp1* block 2 as well as the longer repeat sequences present in other antigens but the algorithm has not been made publically available. However, any algorithm that attempts to calculate the repeat length based on coverage will result in errors.

Chapter 3 - Obtaining a universal catalogue of MSP-1 block 2 epitope sequences from short read data

3.1 Introduction

Predicted intrinsically disordered sequences are abundant in *P. falciparum* proteins and correlate with the presence of repeats (Feng et al., 2006). These disordered sequences are also common in known targets of immune responses and are predicted to present linear B-cell epitopes (Guy et al., 2015, Feng et al., 2006). Antibodies recognising the repetitive linear epitopes present in RTS,S, the only licensed malaria vaccine, correlate with protection in vaccinated individuals and have been shown to prevent infection in a mouse model (Kester et al., 2009, Olotu et al., 2011, Foquet et al., 2014).

The block 2 region of MSP-1 is encoded by a highly polymorphic region of the *msp1* gene that can be classified into four allelic types (section 2.1). The K1-like, MAD20-like and MR recombinant alleles all encode tripeptide repeats which vary both in length and composition. The variation in K1-like repeats is of a higher complexity than that seen for MAD20-like and MR recombinant alleles (section 2.3.9). RO-33-like sequences do not encode repeats and are conserved, with just 6 non-synonymous single nucleotide polymorphisms (SNPs) resulting in 7 different peptide sequences (section 2.3.9). Despite the marked difference in the encoded peptide sequences four allelic types are predicted by both DISOPRED (Jones and Cozzetto, 2015) and DisEMBL (Linding et al., 2003) algorithms to be disordered. Predictions made with BepiPred (Larsen et al., 2006) highlight this region of MSP-1 as the most probable to present linear B-cell epitopes for all three major allelic types (K1-like, MAD20-like and RO-33-like, see section 2.1), consistent with recognition by sera from malaria exposed individuals (section 2.3.10 and (Polley et al., 2003b)). Tetteh *et al* (2005) constructed a synthetic K1-like tripeptide repeat sequence designed to elicit antibody responses to the range of different K1-like repeat structures (section 2.3.9 and (Miller et al., 1993)). The authors of this study first produced 23 peptides corresponding to all combinations of four K1-like tripeptides (serine-alanine-glutamine

(SAQ), serine-glycine-alanine (SGA), serine-glycine-threonine (SGT) and serine-glycine-proline (SGP)) encoded by 49 K1-like *msp1* block 2 sequences from Zambian isolates. They then probed these peptides with 24 adult sera samples from malaria exposed adults that had shown reactivity to K1-like MSP-1 block 2 antigens. This experiment informed the design of a synthetic DNA construct, named the K1 super repeat, encoding 12 of the peptides that showed the best reactivity profile against the panel of sera. This sequence was combined with the RO-33 sequence and the MAD20-like Wellcome sequence to produce a polyvalent hybrid antigen (PVHaf) that was shown to elicit antibodies recognising a range of MSP-1 block 2 sequences, both as recombinant and native antigens, following immunisation of rabbits (Tetteh and Conway, 2011).

Linear B-cell epitopes do not require the tertiary protein structure provided by entire protein domains in order to present biologically relevant epitopes, meaning that multiple epitopes could be efficiently combined into a single polyvalent vaccine. Such a vaccine could be designed to present epitopes from multiple *P. falciparum* proteins as well as multiple sequence variants of a single antigen. This concept has been used to design a polyvalent hybrid antigen protein (PVHaf) that contains a K1-like sequence designed to incorporate a range of K1-like repeat polymorphisms (Tetteh et al., 2005a) fused with the RO-33 and MAD20 sequences (Tetteh and Conway, 2011). The antigen also contains two predicted T-cell epitopes from the block 1 region of MSP-1. This antigen expresses well in *Escherichia coli* (*E. coli*; section 4.2.1) and has been shown to be recognised by sera from clinically immune individuals and elicit antibodies following immunisation of rabbits that recognise both recombinant antigens and parasite strains representing the range of MSP-1 block 2 sequence diversity (Tetteh et al., 2005a, Tetteh and Conway, 2011).

Linear B-cell epitopes typically range in length from 4 to 12 amino acids and can be highly specific, with just a single amino acid substitution resulting in loss of antibody binding (Buus et al., 2012). The successful alignment of repetitive, short read sequences to a library of reference sequences allows for the exploration of potential epitopes without the need for the assembly of the whole sequence,

as even the shortest reads will encode multiple epitopes. The use of short read data to determine epitope frequencies has the potential to inform vaccine design, as hybrid antigens can be designed to incorporate epitopes representing the global sequence diversity. The use of short reads is expedient, as such genomic sequence data are available for large numbers of parasite isolates recently sampled from endemic regions.

This chapter details the analysis of amino acid sequences determined by alignment of short reads to the library of reference sequences (appendix 7.4) described in the previous chapter (section 2.3.4) and their subsequent use in designing polyvalent hybrid antigen constructs. These designs are then tested *in silico* to determine the number of potential epitopes in the set of *de novo* assembled Pf3k *msp1* block 2 sequences (tSRA, section 2.3.9) that are present in each design thus allowing comparison between designs and with the previously described PVHaf (Tetteh and Conway, 2011).

3.2 Materials and methods

3.2.1 Data sources

In order to create short read sequencing data from known sequences, long read sequences were randomly split into “synthetic reads”. These synthetic reads were created from 964 MSP-1 block 2 sequences in the long read dataset (LRD, described in section 2.3.1, additional data file “long_read_sequences.fa”) using a modified version of `to_perfect_reads` (available from <https://github.com/sanger-pathogens/Fastaq>) as described above (section 2.2.2). Raw reads were extracted from BAM files downloaded from the Pf3k project and aligned to a library of MSP-1 block 2 reference sequences as described above (section 2.2.5). MSP-1 block 2 sequences from the Pf3k data were obtained by *de novo* assembly of reads mapping to the MSP-1 locus and reads aligned to the *msp1b2RefLib* as described above (sections 2.2.4 and 2.2.5). The amino acid sequence of MSP-1 block 2 polyvalent hybrid antigen F was obtained from the publication (Tetteh and Conway, 2011).

3.2.2 Translation of aligned reads

Reads were first aligned to a *msp1b2RefLib* of 15 MSP-1 block 2 sequences (appendix 7.4) as described above (section 2.2.5). The resulting BAM files containing all mapped reads were converted into SAM format using SAMtools (Li et al., 2009). Python scripts (appendix 7.7) were developed to extract the cigar string, containing mapping information, and obtain the position of the first base of the read relative to the sequence it is mapped to. This position was then used to translate the read in frame with the sequence to which it was aligned.

3.2.3 Analysis of nonamers and design of minimal polyvalent antigens

Python functions (appendix 7.7) were defined to split translated reads into all possible nonamer sequences and output all nonamer sequences along with the fraction of all reads in which the nonamer occurred for each sample. Python was also used to write a function to determine the population-wide frequency, scaled to account for inter sample variation in coverage depth (appendix 7.5), for each unique nonamer, calculated by summing, across all samples containing the nonamer, the number of times each nonamer occurred in a given sample divided by the total number of aligned reads in that sample (figure 3.2). The python function “nonamerise” was also used to split amino acid sequences encoded by long read or *de novo* assembled sequences into nonamer sequences for comparison and validation.

3.3 Results

3.3.1 Short reads can be accurately translated based on alignment to a sequence library

To determine whether DNA sequence reads could be translated to amino acid sequence based on alignment to a sequence library, synthetic reads generated from long read dataset (LRD, section 2.3.1) sequences (section 2.2.2) were first aligned to a sequence library and then translated. The whole of the read spanning the block 2 region was translated using the mapping information to determine the correct frame. The resulting reads were then split into nonamer sequences, as the

standard, linear antibody epitope is nine amino acids long (Buus et al., 2012). The nonamers were then listed and duplicates removed. By comparing the unique nonamers obtained from the translation of synthetic reads to those determined by splitting all LRD sequences (section 2.3.1) used for synthetic read generation into nonamers it was determined that 84.9% (1155/1361) and 84.3% (1147/1361) of unique nonamers were recovered by translation of aligned 100 bp and 75 bp synthetic short read sequences respectively. The nonamers that were missing following translation of aligned reads resulted from rare sequence variants not represented in the *msp1b2RefLib*, meaning that the synthetic reads containing these sequences did not map. No new nonamers were generated, indicating accurate translation of aligned reads.

3.3.2 Prevalence of nonamer epitopes varies by region

Following the success of translating synthetic reads (see above section 3.3.1), all 9.39×10^5 reads from the Pf3k dataset that had been aligned (section 2.2.5) to the *msp1b2RefLib* (appendix 7.4) were translated as described above (sections 3.2.2 and 3.3.1) and broken into all possible nonamer sequences resulting in a total of 1.47×10^7 nonamers. By comparing nonamer sequences to MSP-1 block 2 sequences encoded by sequences in the LRD and short read assembled (SRA) datasets it was determined that 8.06×10^6 (54.8%) nonamer sequences were of the K1 type, 2.99×10^6 (20.3%) were of the MAD20 type and 1.46×10^6 (9.92%) were RO-33-like. Whilst K1-like and MAD20-like sequences were detected in almost equal proportions globally (table 2.2), the increased coverage of African and culture adapted samples (appendix 7.5), both of which are enriched for K1-type alleles, combined with the increased length of K1-like alleles results in the increased fraction of nonamer sequences being identified as K1-like. Due to the similarity between the MR recombinant 5' sequence and MAD20-like 5' sequence and between the MR recombinant 3' sequence and 3' RO-33-like sequences, some of the nonamers identified as MAD20-like or RO-33-like would have originated from MR recombinant sequences. The unique sequence generated by recombination of MAD20-like and RO-33 like sequence present in MR alleles was identified in 3.07×10^4 nonamers. There were

2.17×10^6 (14.7% of total) nonamer sequences that were not present in any sequences present in either the LRD or SRA datasets. These nonamers are the result of sequencing errors and frameshifts introduced by gapped mapping of reads and each sequence only occurs once. These sequences were retained for further analysis as application of this method to antigens for which there is not such a wealth of sequence data would not permit the filtering of nonamer sequences by comparison to known sequences.

Removing duplicate nonamer sequences left a total of 1.99×10^5 unique sequences globally. The majority (99.6%) of these unique sequences could not be identified by comparison to known MSP-1 block 2 sequences. This is due to the fact that sequencing errors and frameshifts introduced by gapped alignments result in the generation of a large number of unique sequences. Of the 764 unique sequences that could be matched to nonamer sequences of the known K1-like, MAD20-like or RO-33-like sequences present in LRD and SRA databases, 372 (48.7%) were of the K1-type, 229 (30.0%) of the MAD20-type and 163 (21.3%) of the RO-33 type. This is expected given the increased complexity of K1-like tripeptide repeat sequences compared to those encoded by MAD20-like alleles, and the lack of repeat sequences in RO-33-like alleles resulting in far less polymorphism (section 2.3.9).

African samples yielded a total of 753 unique nonamer sequences that could be identified, of which 372 (49.4%) were of the K1 type, 218 (29.0%) were of the MAD20-type and 163 (21.6%) were of the RO-33 type. One hundred and fifty of these nonamers sequences were only found in African samples, with 98 (65.3%) of these belonging to the K1 type, 10 (6.67%) belonging to the MAD20 type and 42 (28.0%) belonging to the RO-33 type. Of the 753 unique nonamer sequences found in Africa that could be classified as K1-like, MAD20-like or RO-33-like, 603 (80%) were also found in Asia where an additional 11 sequences were found that are not seen in Africa, resulting in 614 unique nonamer sequences in Asia. There were 274 K1 type unique nonamer sequences present in Asia, all of which were also seen in Africa. The K1 type nonamer sequences account for a slightly smaller

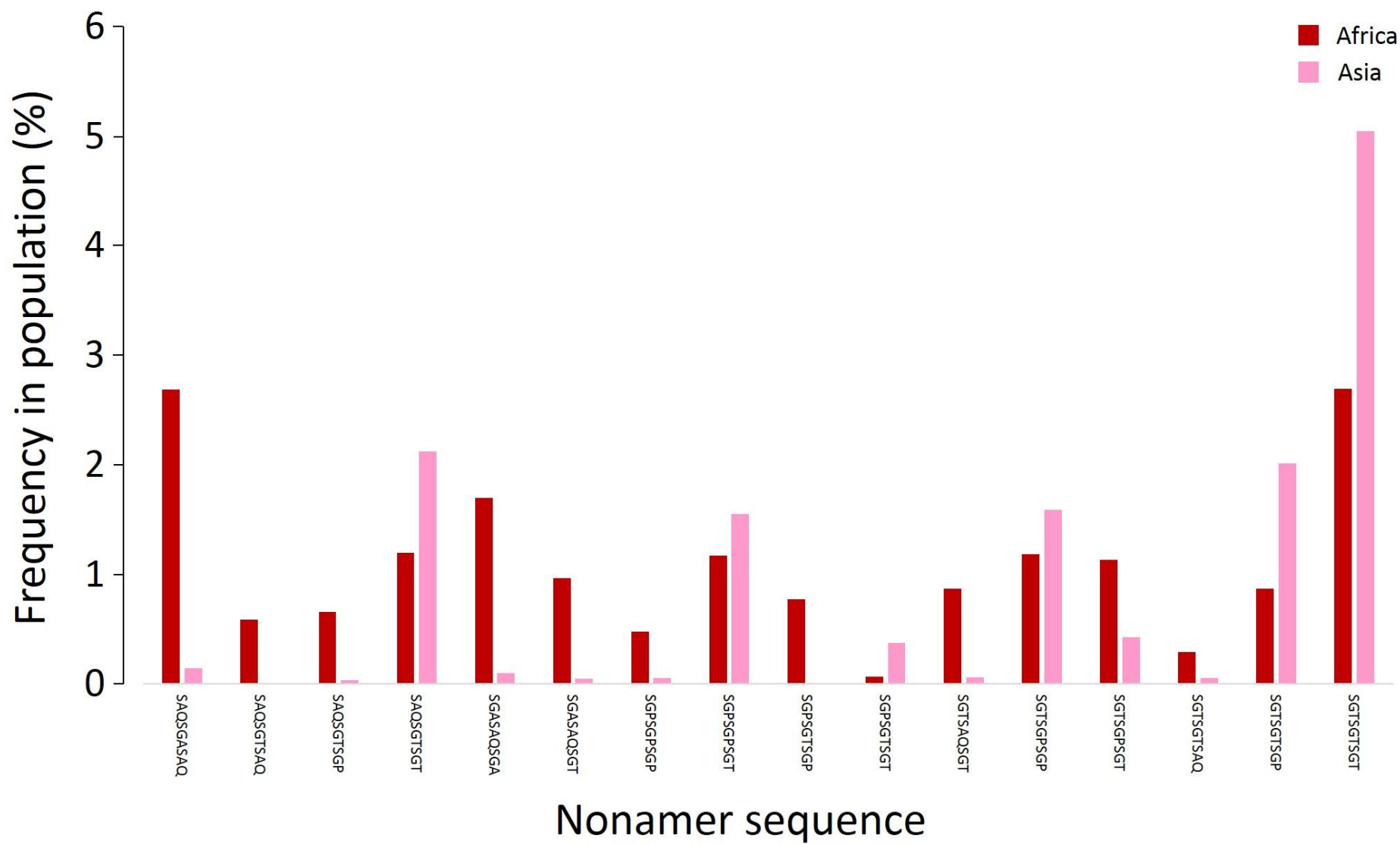
proportion (44.6%) of unique nonamer sequences in Asia than in Africa but this is not significant ($p = 0.088$, χ^2 test), due to the decreased complexity of K1-like alleles in Asia (section 2.3.9), but still comprise the greatest proportion of unique nonamer sequences detected in Asia. There were a similar number of MAD20 type nonamer sequences that were exclusive to Asia (11) or Africa (10). The small increase in complexity of Asian MAD20-like sequences relative to African MAD20-like sequences (due to the presence of serine-valine-alanine-serine-valine-alanine (SVASVA) motifs and the serine-aspartic acid-glycine (SDG) tripeptide) results in a small but significant increase in the proportion of unique nonamers of the MAD20 type which represent 219 (35.7%) of all unique nonamer sequences ($p < 0.001$, χ^2 test). Due to the fact that the SNP in the RO-33 allele that is exclusive to Asia is also found in an African allele (figure 2.14), all Asian RO-33-like nonamers are also found in Africa. The slight reduction, relative to Africa, in the proportion of unique RO-33-like nonamers in Asia, which account for 121 (19.7%) of unique nonamers found in Asia, is a result of the reduced number of RO-33-like alleles present in Asia compared to Africa (section 2.3.9), although this shift is too small to be statistically significant ($p = 0.38$, χ^2 test).

Due to the difference in coverage depth between samples (appendix 7.5), it was necessary to scale the number of nonamer sequences counted for each sample against the number of reads mapped reads (section 3.2.3, figure 3.2). As the frequencies of different allelic types vary between continents (table 2.2) the frequency of each nonamer was determined as a percentage of all nonamers of that allelic type present in either Africa or Asia. The majority of tripeptide nonamer sequences occur at low frequencies. Due to the fact that the vast majority of variation within K1-like and MAD20-like allelic families is due to changes in the tripeptide repeat, nonamers arising from this region were examined and, for clarity, only nonamers beginning with the first serine of the tripeptide were considered. Five K1 type tripeptide nonamers are found at high ($> 1.0\%$) frequency in Asia (figure 3.1 a), as they are all encoded by the K1-like repeat structure that is found in over 80% of Asian K1-like alleles (figure 2.13 and section 2.3.9). These same nonamers are also found at high, but slightly reduced, frequencies in Africa. There are two nonamers that are above 1% frequency in Africa but

rare in Asia (figure 3.1 a) both arising from the SAQSGA motif encoded by K1-like alleles of the 3D7-like subtype, which is at a far higher frequency in African isolates (section 2.3.9). An additional nonamer sequence, SGTSGPSGT, is present at high (>0.4%) frequency in both Africa and Asia. This nonamer is found in just over a fifth of K1-like sequences with the Asian or K1-like subtype structure in Asia, but is present in a greater number of alleles in Africa as it occurs in a wider array of different repeat structures encoded by African alleles. The remaining African nonamers present at high (>0.25%) frequencies are rare in Asia, as they arise from repeat structures that are not commonly found in Asia (section 2.3.9).

The picture for MAD20-like tripeptide nonamers is the reverse of that seen for K1-like, with five nonamers present in the repeat structure which is most common in Africa at high frequencies in both continents but notably higher in Africa (figure 3.1 b). As predicted from analysis of the differences in MAD20-like repeat structure, the nonamer derived from expanded SGG repeats (SGGSGGSGG) is higher in Africa than in Asia. Five nonamers originating from Asian repeat structures containing SVASVA motifs or SSG tripeptides are seen at high frequency in Asia but are at low frequency or absent from African isolates. It is interesting to note however, that the SVASVA motif, which was not seen in *de novo* assembled, African MAD20-like tripeptide repeats, is present in at low frequencies in Africa. Nonamers containing the SKGSVA and SKGSVT motifs are common in both Africa and Asia, with SKGSVT being at a higher frequency in Africa, as would be predicted from the *de novo* assembled sequences which show this motif to be enriched in Africa (section 2.3.9).

a



b

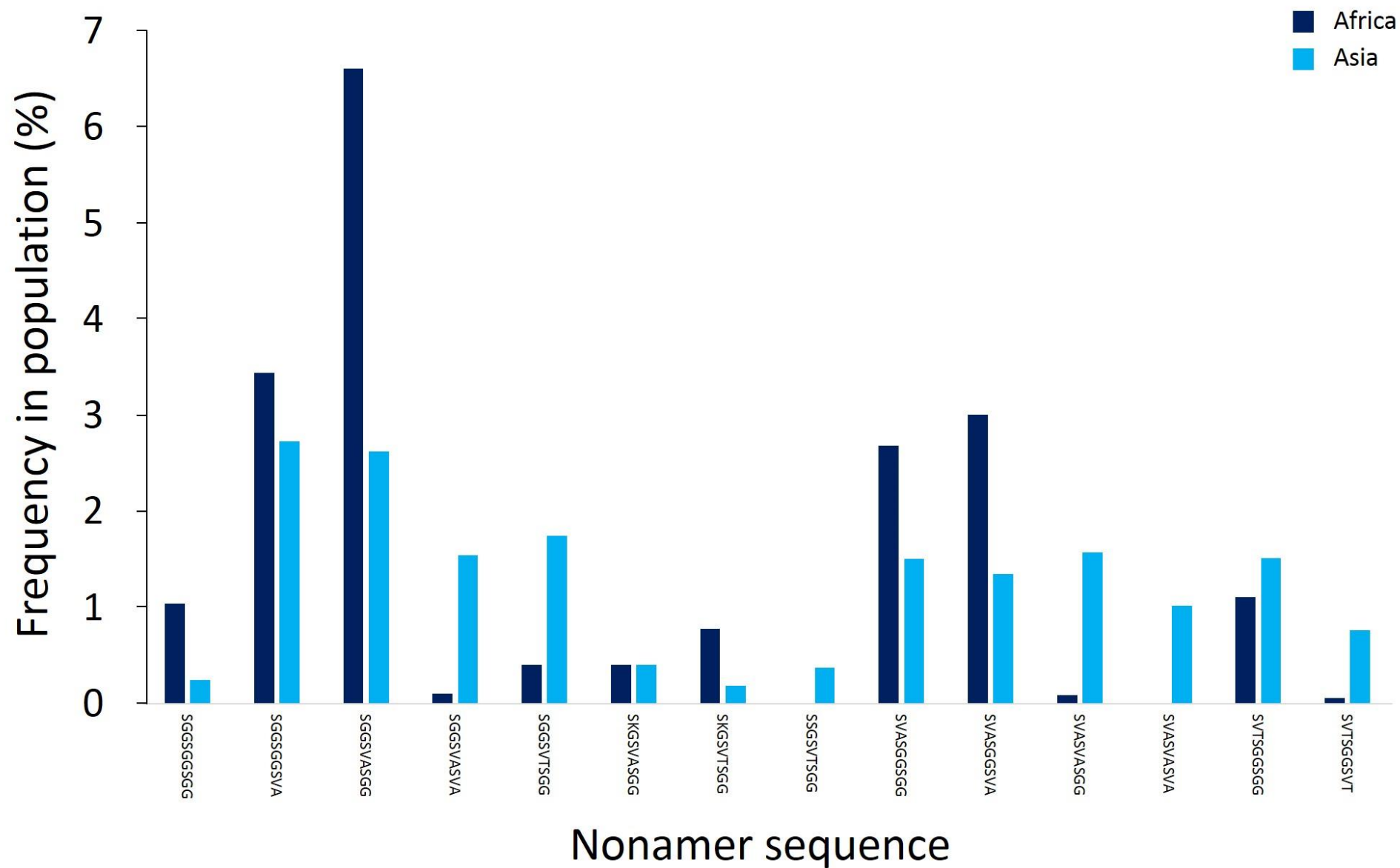


Figure 3.1. Regional variation in frequencies of most common nonamers. 75-100 bp reads were aligned to a library of MSP-1 block 2 reference sequences and then translated (section 3.2.2). The translated reads were then broken into nonamer sequences and the frequency of each nonamer, scaled for coverage depth (figure 3.2), was calculated. The frequency of nonamers originating from K1-like tripeptide repeats (a) are shown as the percentage of all K1-like nonamers present in either in Africa (dark red) or Asia (pink). The frequency of MAD20-like tripeptide nonamers (b) are shown as a percentage of all MAD20-like nonamers in Africa (dark blue) or Asia (light blue). For clarity, only nonamers beginning with the first serine of the tripeptide that have a frequency of > 0.25% are shown.

3.3.3 An algorithm for designing polyvalent hybrid antigens was designed and optimised

An algorithm was developed in order to use the translated reads to design peptide sequences that would include as many nonamer epitopes presented by the different MSP-1 block 2 alleles as possible in a minimal length by combining nonamer sequences derived from translated reads (figure 3.2). The algorithm considers two factors when deciding whether or not to include a given nonamer sequence. The first is the frequency of the nonamer in the population, which is determined by summing the frequency of the nonamer in each sample after normalising for differences of coverage depth between samples (figure 3.2). The second factor is the number of residues that need to be added to incorporate a given nonamer into the proposed antigen sequence such that a minimal length is achieved. These two factors are combined in an inclusion score (figure 3.2) which is calculated for each nonamer sequence with every iteration of the algorithm. The seed forming propensity (sfp) is a linear scaling factor that determines the degree to which the algorithm will incorporate nonamers based on their relative frequency as opposed to the number of residues that need to be added to the sequence; the higher the sfp the greater weight that is given to the frequency of the nonamers and the lower the sfp the greater weight that is given to the ability to incorporate a given nonamer with a minimal increase in antigen length.

$$z = l + (sfp \times f_s)$$

Equation 3.1. Formula for calculating nonamer inclusion scores. Inclusion scores (z) for each nonamer are calculated in each iteration as the sum of the length of the overlap (l) and the product of the seed forming propensity parameter (sfp) and the scaled frequency (f_s).

The algorithm will use the nonamer with the highest frequency in the population as the first seed.

Inclusion scores are then calculated for all remaining nonamers and the nonamer with the highest score is then added to the antigen. If there is an overlap between the new nonamer and the beginning or end of any seed, the longest overlapping sequence will be used to incorporate the new nonamer into a seed sequence; if the new nonamer has no overlap with any seeds it will be added as an additional seed sequence. The algorithm then checks whether the addition of the new nonamer has created any overlaps between the ends of any seed sequences and, if so, uses the shared sequence to combine the pair of overlapping sequences into one seed sequence. This process is performed iteratively until a pre-set maximum length for the antigen is reached.

In order to optimise the sfp parameter, synthetic reads were created from *mvp1* block 2 sequences from the LRD (described in section 2.3.1) and these reads were mapped (section 2.2.5) to the *mvp1b2RefLib* (appendix 7.4) and translated (section 3.2.2). The unique nonamers generated from the translated reads were then fed into the polyvalent hybrid antigen design algorithm with varying sfp parameter values and a maximum length of 231 amino acids (set to be the same length as the MSP-1 block 2 sequences present in the manually designed polyvalent hybrid antigen F (Tetteh and Conway, 2011)). The number of nonamer epitopes of each LRD sequence that was present in the antigen was determined and the percentage of these epitopes as the total of all unique epitopes was calculated for polyvalent hybrid antigen sequences generated by the algorithm with a range (1-200) of sfp parameter values (figure 3.3). Altering the sfp parameter resulted in significant changes in the coverage of MSP-1 block 2 nonamers by the resulting antigen sequence, with median values ranging from 64 to 93%. The highest median coverage of all allelic types combined was for the antigen

design produced by the algorithm with an sfp of 51. However, an sfp of 71 was chosen as optimal as the antigen produced with this sfp parameter value, whilst having a lower median coverage of K1-like alleles, had better median coverage of other allelic types (table 3.1).

The antigen produced by the algorithm (with an optimal sfp of 71) was compared to the polyvalent hybrid antigen designed by eye. Across all sequences in the LRD, the median percentage of nonamer epitopes present in polyvalent hybrid antigen F was determined to be 76%, significantly lower than that for the antigen produced by the optimised algorithm (92%, $p < 0.001$, Wilcoxon rank-sum). The K1-like super repeat (Tetteh et al., 2005b), which is present in the polyvalent hybrid antigen F, contains a high proportion of K1 type nonamer sequences found in the LRD (median 93% across all K1-like sequences) with at least 40% of nonamers in each LRD K1-like sequence being present in the polyvalent hybrid antigen F and the majority having over 90% of their nonamer sequences represented. The antigen designed using nonamers of translated synthetic reads by the algorithm with an optimised sfp of 71 made only a very slight improvement on this, with a median value of 94% of nonamers present in LRD sequences also present in the algorithm and with at least 45% of nonamers in each sequence represented (table 3.1). In addition to the K1-super repeat, the polyvalent hybrid antigen contains the Wellcome MAD20-like allele and RO-33 sequences. The antigen designed by the algorithm using nonamer sequences from the LRD included a wider range of MAD20-like motifs as well as an MR-like sequence, with the G97D mutation (section 2.3.9). The inclusion of these additional sequences means that the antigen sequence design by the algorithm performed better than the polyvalent hybrid antigen F in regards to RO-33-like and MR recombinant sequences in the LRD and much better in regards to MAD20-like sequences (table 3.1). In order to determine whether equivalent coverage of nonamers could be achieved with a shorter antigen sequence, the same approach was used to analyse antigens designed with a range (50-225 amino acid) lengths. Decreasing the antigen length dramatically reduces the number of sequences with moderate (> 50%) and high (>75%) of their composite nonamer sequences included in the antigen. It was therefore decided to continue with an antigen length of 231 amino acids.

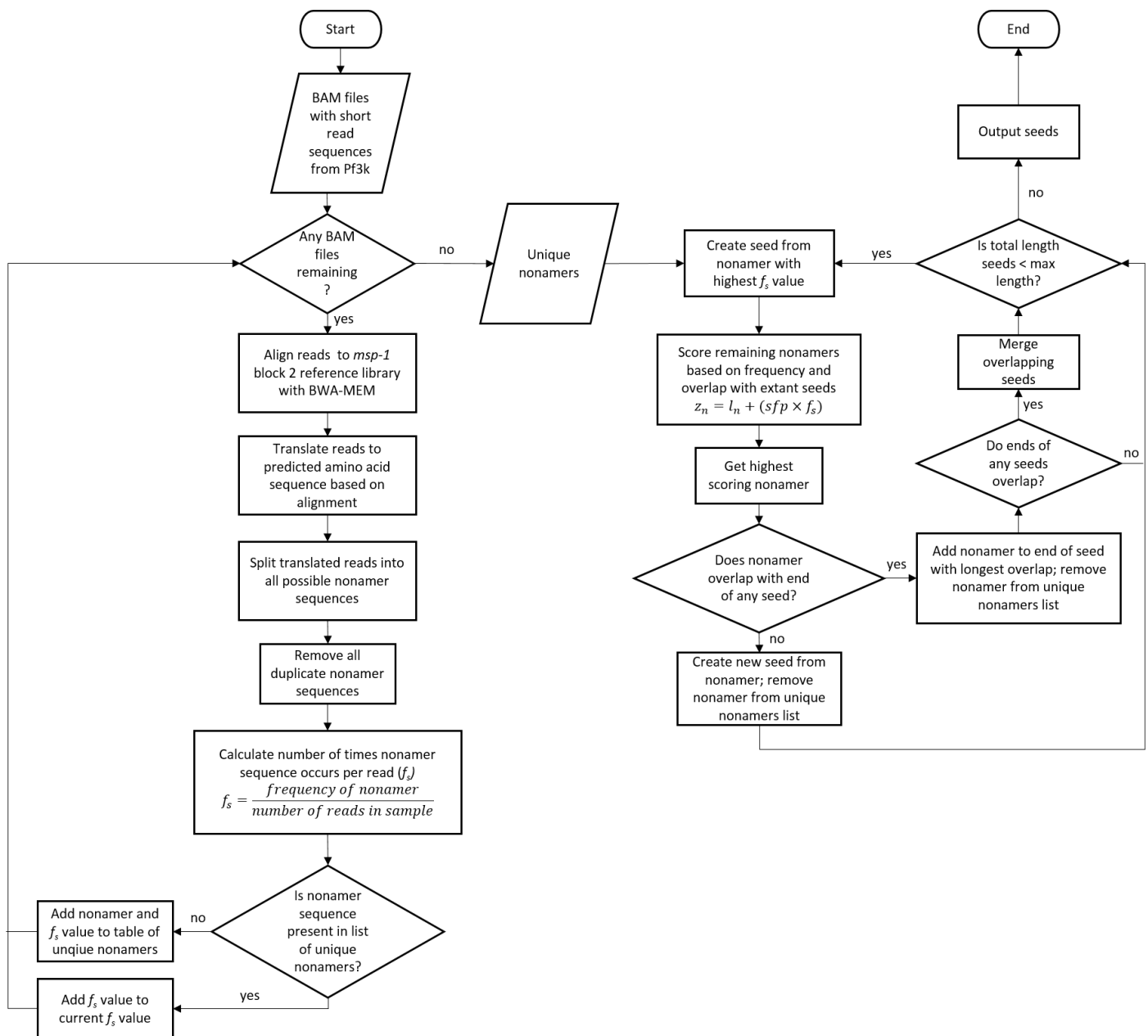


Figure 3.2 Flow chart showing how polyvalent hybrid antigens were generated from short sequence reads.

Short sequence reads are first aligned to the *msp1b2RefLib* of *msp1* block 2 sequences (see Chapter 2). Aligned reads are then translated (see above section 3.2.2) and the unique nonamer amino acid sequences stored. Scaled frequencies are calculated for each unique nonamer based on the number of reads in which it occurs and the total number of reads mapping to the *msp1b2RefLib* for the sample. The scaled frequencies are summed for each unique nonamer sequence and the nonamer with the highest scaled frequency is selected as the first seed. The remaining nonamers are then scored based on the product of their scaled frequency (multiplied by a linear scaling factor) and the degree of overlap with the end of the seed(s) (overlaps of less than three amino acid residues are ignored). The highest scoring nonamer is then added to the seed(s) by either adding amino acids to the end of an extant seed or by adding the nonamer as a new seed. After adding a nonamer to a seed, overlaps between seeds are found and, if present, seeds are merged. This process is repeated iteratively until a predetermined maximum length has been reached for the antigen.

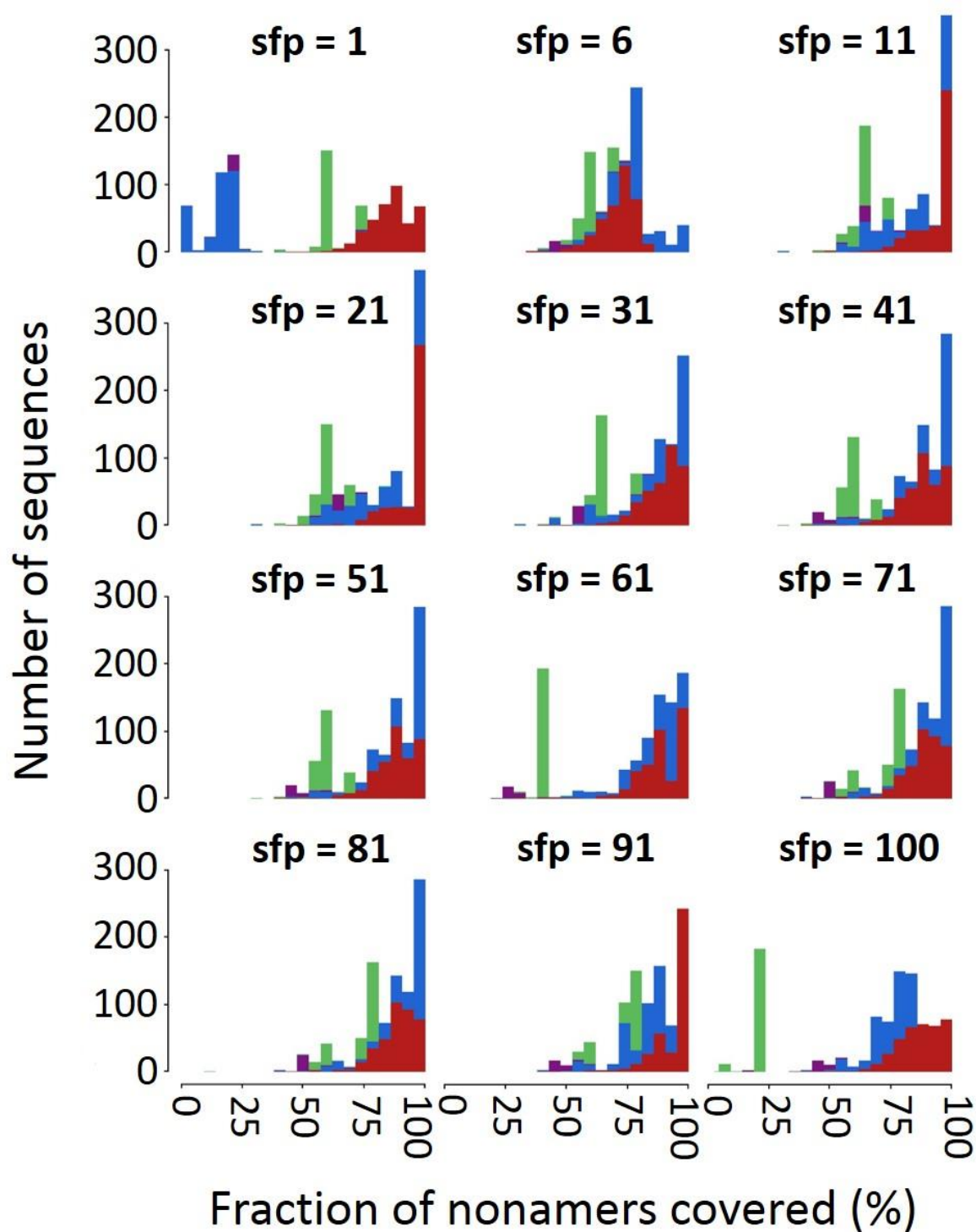


Figure 3.3 Optimisation of seed forming propensity for polyvalent hybrid antigen algorithm. Nonamers generated from translation (section 3.2.2) of synthetic reads (section 2.2.2) created from sequences deposited in LRD (section 2.3.1) and aligned to the *msp1b2*RefLib (appendix 7.4) were used to create polyvalent hybrid antigen sequences with the algorithm described above using a range of seed forming propensity values (sfp; see above). Histograms show the distribution of the percentage of all nonamers represented in the resultant polyvalent hybrid antigen sequences for each sequence in the LRD. Bars are coloured by the allelic type of the sequence with K1-like (red), MAD20-like (blue), RO-33-like (green) and MR recombinant (purple) sequences shown.

spf	K1-like	MAD20-like	RO-33-like	MR	All
1	0.91 (0.49-1.0)	0.19 (0.0-0.75)	0.64 (0.42-0.76)	0.2 (0.14-0.78)	0.64 (0.0-1.0)
2	0.91 (0.49-1.0)	0.19 (0.0-0.75)	0.64 (0.42-0.76)	0.2 (0.14-0.78)	0.64 (0.0-1.0)
4	0.91 (0.45-1.0)	0.76 (0.4-1.0)	0.62 (0.42-0.73)	0.47 (0.47-0.78)	0.8 (0.4-1.0)
6	0.76 (0.38-0.86)	0.81 (0.42-1.0)	0.62 (0.42-0.76)	0.47 (0.47-0.78)	0.76 (0.38-1.0)
8	1.0 (0.45-1.0)	0.81 (0.42-1.0)	0.6 (0.4-0.71)	0.47 (0.47-0.78)	0.82 (0.4-1.0)
11	1.0 (0.47-1.0)	0.9 (0.33-1.0)	0.69 (0.49-0.78)	0.65 (0.57-0.84)	0.9 (0.33-1.0)
21	1.0 (0.48-1.0)	0.9 (0.33-1.0)	0.6 (0.4-0.71)	0.65 (0.57-0.78)	0.93 (0.33-1.0)
31	0.96 (0.45-1.0)	0.94 (0.33-1.0)	0.67 (0.42-0.82)	0.55 (0.55-0.86)	0.91 (0.33-1.0)
41	0.94 (0.45-1.0)	1.0 (0.49-1.0)	0.62 (0.33-0.73)	0.49 (0.49-0.69)	0.92 (0.33-1.0)
51	0.98 (0.47-1.0)	0.96 (0.42-1.0)	0.62 (0.33-0.73)	0.49 (0.49-0.72)	0.93 (0.33-1.0)
61	0.94 (0.31-1.0)	0.94 (0.51-1.0)	0.44 (0.07-0.44)	0.29 (0.24-0.45)	0.9 (0.07-1.0)
71	0.94 (0.45-1.0)	1.0 (0.44-1.0)	0.8 (0.13-0.8)	0.51 (0.5-0.74)	0.92 (0.13-1.0)
81	0.94 (0.45-1.0)	1.0 (0.44-1.0)	0.8 (0.13-0.8)	0.51 (0.5-0.74)	0.92 (0.13-1.0)
91	1.0 (0.45-1.0)	0.88 (0.44-0.98)	0.8 (0.13-0.8)	0.49 (0.45-0.61)	0.89 (0.13-1.0)
100	0.93 (0.38-1.0)	0.8 (0.42-0.89)	0.22 (0.0-0.23)	0.49 (0.18-0.59)	0.8 (0.0-1.0)
110	0.92 (0.42-1.0)	0.8 (0.42-1.0)	0.27 (0.06-0.28)	0.39 (0.22-0.48)	0.8 (0.06-1.0)
120	0.96 (0.4-1.0)	0.8 (0.42-1.0)	0.24 (0.0-0.26)	0.49 (0.2-0.57)	0.8 (0.0-1.0)
130	0.94 (0.44-1.0)	0.74 (0.33-0.91)	0.29 (0.0-0.3)	0.47 (0.24-0.59)	0.8 (0.0-1.0)
140	0.94 (0.45-1.0)	0.74 (0.33-0.91)	0.31 (0.0-0.33)	0.45 (0.25-0.57)	0.8 (0.0-1.0)
150	0.94 (0.45-1.0)	0.74 (0.33-0.91)	0.31 (0.0-0.33)	0.47 (0.25-0.59)	0.8 (0.0-1.0)
160	0.94 (0.45-1.0)	0.74 (0.33-0.91)	0.31 (0.0-0.33)	0.47 (0.25-0.59)	0.8 (0.0-1.0)
170	1.0 (0.45-1.0)	0.8 (0.42-0.89)	0.31 (0.0-0.33)	0.47 (0.25-0.57)	0.8 (0.0-1.0)
180	1.0 (0.45-1.0)	0.8 (0.42-0.89)	0.31 (0.0-0.33)	0.47 (0.25-0.57)	0.8 (0.0-1.0)
190	0.94 (0.42-1.0)	0.74 (0.33-0.91)	0.36 (0.07-0.37)	0.37 (0.27-0.5)	0.8 (0.07-1.0)
200	0.94 (0.42-1.0)	0.74 (0.33-0.91)	0.36 (0.07-0.37)	0.37 (0.27-0.5)	0.8 (0.07-1.0)
PVH					
antigen F	0.93 (0.4-0.96)	0.65 (0.29-0.97)	0.76 (0.16-0.76)	0.45 (0.36-0.63)	0.76 (0.16-0.97)

Table 3.1. Coverage of nonamer epitopes in LRD sequences by antigens designed by algorithm using translated reads. Synthetic reads were generated from *msh1* block 2 sequences in LRD (section 2.3.1, additional data file “long_read_sequences.fa”). These reads were then aligned and translated, as described above. The nonamer sequences present in these reads were used to generate polyvalent hybrid antigen sequences using the algorithm described above (section 3.3.3) with a range of seed forming propensity (sfp) parameters. The fraction of the unique nonamers present in each LRD sequence was then determined for every antigen generated. Fractions of nonamers covered for different allelic types are shown for each antigen with the range of coverage in parentheses. The antigen resulting from the optimal spf value is highlighted in bold. The polyvalent hybrid (PVH) antigen F, designed by eye, is shown for comparison.

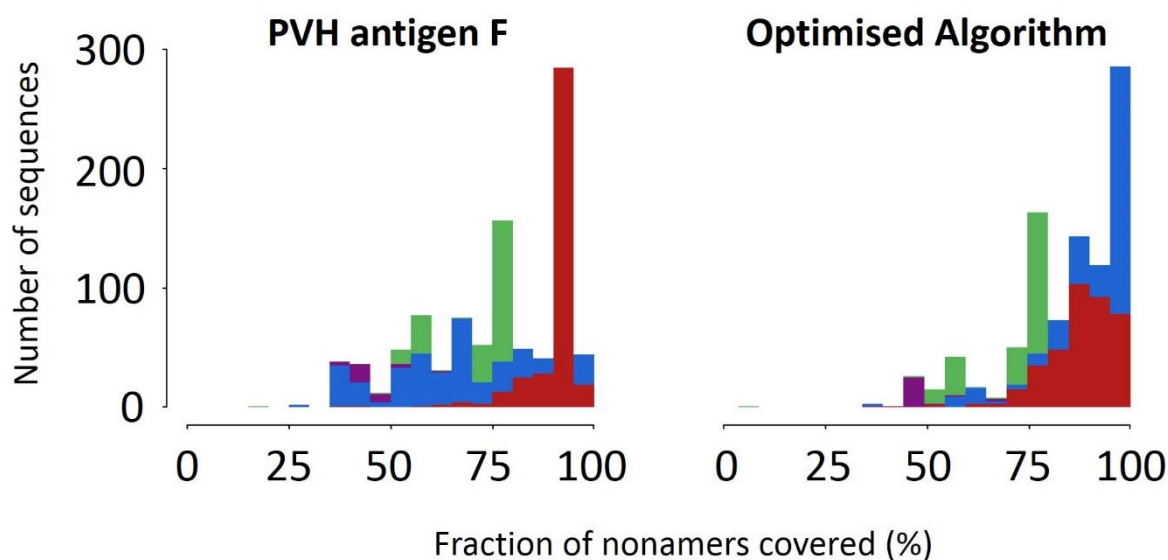


Figure 3.4 *In silico* comparison of antigens designed by algorithm and by eye. Synthetic reads created from LRD sequences (section 2.3.1) were translated and used to create a polyvalent hybrid antigen with an algorithm that combines the most frequent nonamer sequences (section 3.2.3). The antigen produced following optimisation of this algorithm (see above) was then compared to the polyvalent hybrid antigen F designed manually (Tetteh and Conway, 2011); the number of polyvalent hybrid antigen nonamer sequences found in each MSP-1 block 2 sequence from LRD as a fraction of all unique nonamers in that sequence was determined for each antigen. Histograms show the distribution of the percentage of unique nonamers covered by the polyvalent hybrid (PVH) antigen F (left) and the antigen produced by the optimised algorithm (right), with colours showing the allelic type of the sequence; K1 (red), MAD20 (blue), RO-33 (green) and MR (purple)

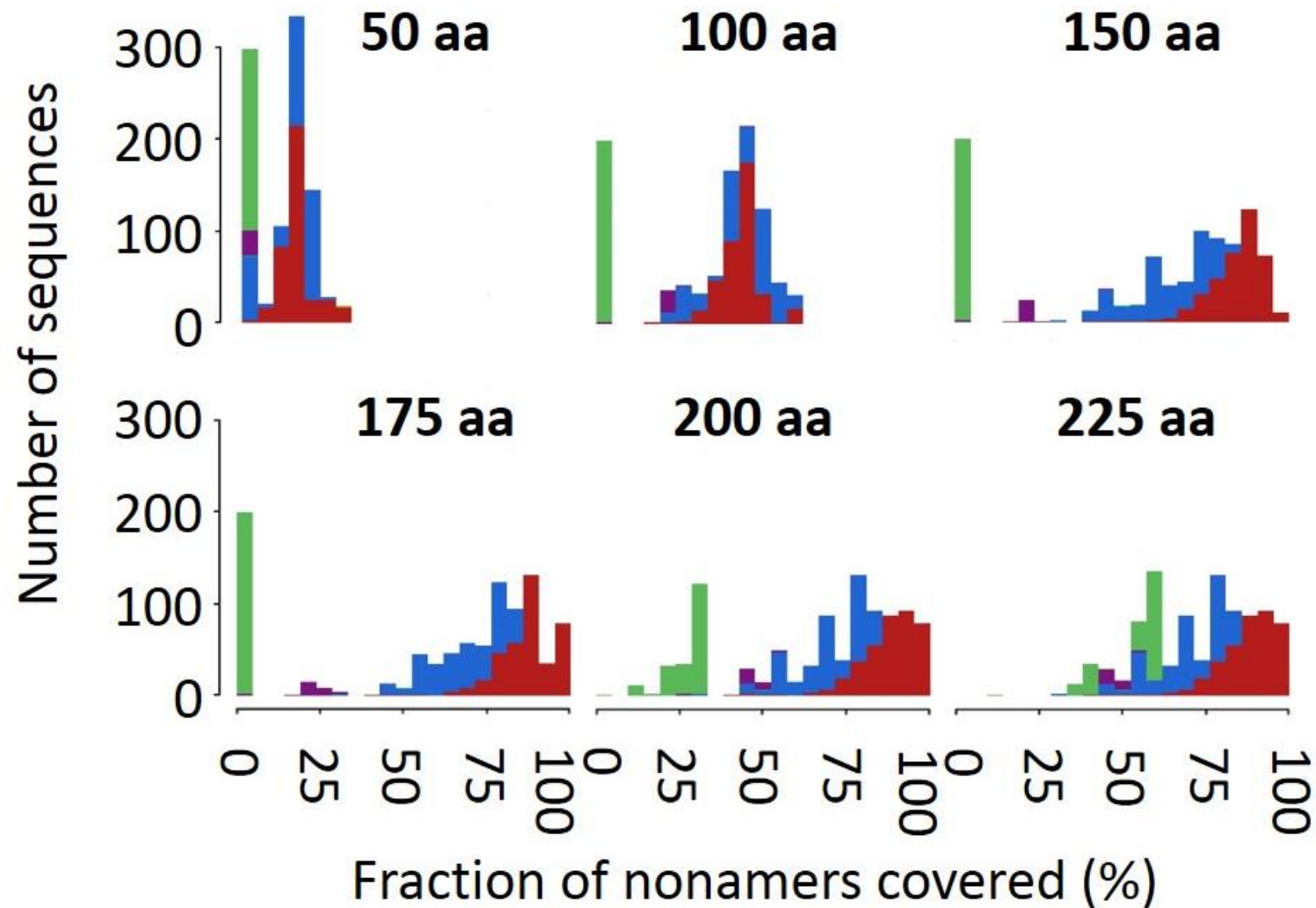


Figure 3.5 Effect of length of antigen on coverage of sequences. The algorithm for designing polyvalent hybrid antigens was used to generate antigens with a range (50 – 225 amino acids) of lengths from nonamer amino acid sequences extracted from synthetic reads (section 2.2.2) aligned to the *msp1b2RefLib* and then translated (section 3.2.2). The resultant antigens were then compared to all sequences from LRD (section 2.3.1) to determine the fraction of nonamer epitopes presented in each sequence that were covered by each antigen. Histograms show the frequency distribution of the percentage of epitopes covered by the antigens of different length for the three major allelic types; K1 (red), MAD20 (blue), RO-33 (green) and for the MR-like recombinant alleles (purple)

3.3.4. Design of region specific polyvalent antigens to incorporate range of epitopes

The optimised algorithm (section 3.3.3) was then used to design polyvalent hybrid antigens based on the frequency of nonamer sequences obtained from aligning reads in the Pf3k dataset to the *m*sp1b2RefLib (appendix 7.4) and translating these reads according to their alignment (section 3.2.2). The maximum length set for the antigens designed by the algorithm was chosen to be 231 amino acids, as this is the length of the *m*sp1 block 2 sequences present in the previously designed polyvalent hybrid antigen which is easily expressed in *E. coli* ((Tetteh and Conway, 2011) and section 4.2.1). Due to the variation in nonamer frequencies between samples from Africa and Asia (figure 3.1), in addition to designing a global antigen sequence based on the nonamer frequencies seen across all samples, antigen sequences were also designed based only on the frequency of nonamers present in either Africa or Asia. For each of the three sample sets the algorithm output a single RO-33-like sequence. When using the global and African samples sets, one complete and one incomplete K1-like and one complete and one incomplete MAD20-like sequence. When using the Asian samples, the algorithm out put one complete K1-like sequence and five incomplete MAD20-like sequences. The tripeptide repeat sequences from incomplete K1- and MAD20-like sequences produced by the algorithm were then concatenated by eye to give one complete K1-like sequence and one complete MAD20-like sequence for each antigen design (figure 3.6). Whilst all three antigen designs contain sequences from the three major allelic types, it is clear that a greater proportion of the African polyvalent hybrid antigen comprises K1-like sequence variants, whereas the Asian polyvalent hybrid antigen contains a greater proportion of MAD20-like sequence. This is caused by the variation in the frequency of allelic types between Africa and Asia (table 2.2) but also by the increased complexity of K1-like repeat structures in Africa compared to Asia and the increased complexity of MAD20-like repeat structures in Asia compared to Africa. Increased complexity of repeat structure leads to a greater number of unique nonamer sequences and, therefore, the need to create a longer synthetic repeat sequence to represent these nonamers.

The sequences obtained by *de novo* assembly of Pf3k data were then used to test the coverage of the nonamer peptide sequences in proposed polyvalent hybrid antigens. As described previously (section 3.3.3), the antigens designed by the algorithm were tested against each sequence present in the SRA dataset individually; the number of nonamers in a given SRA sequence that were present in the antigen sequence was determined as a percentage of all nonamers in the given sequence. This gives a percentage coverage for the antigen for each of the 1523 sequences in the SRA dataset. Each antigen could then be assessed by distribution of coverage achieved across the sequence in the SRA (figure 3.7). Results of this analysis show that the global antigen achieved a median coverage of 85% (range 47-100%) of nonamers (table 3.2). This was slightly lower than the performance of the previously described polyvalent hybrid antigen, (PVHAF; 93%, range 33-100%) but this difference was not significant ($p = 0.24$, Wilcoxon rank-sum). Comparing the coverage of nonamers in SRA sequences of each allelic type by these two antigens shows that the PVHAF performed better against K1-like and RO-33-like sequences but the novel PVHAG performed better against MAD20-like sequences ($p < 0.001$, Wilcoxon rank-sum, figure 3.7). This is to be expected as the PVHAF includes the K1 super repeat, which is designed to incorporate a range of K1-like nonamers (Tetteh et al., 2005b), but only a single MAD20-like allele (Tetteh and Conway, 2011). The reason that the PVHAG achieves a poorer coverage of RO-33-like sequences is that the variation at position 67, where an aspartic acid is seen in almost all African alleles and a glycine is seen in the vast majority of Asian alleles, is not included in the sequence produced by the algorithm, meaning that nonamers containing this position are not covered. This shows how this approach to antigen design can work well for polymorphic repeat sequences, but is not ideal for dealing with dimorphism at a single amino acid position.

As expected, both the African (PVHAaf) and Asian (PVHAas) antigens contained a higher proportion of sequences from the populations that they were designed for (table 3.2, figure 3.7). Whilst the improvement of the Asian antigen over PVHAF was starker, both antigens cover a higher percentage of nonamers present in sequences from their respective continents than PVHAF ($p < 0.001$, Wilcoxon

rank-sum, table 3.2, figure 3.7). This is due to the fact that the antigens designed for either Africa or Asia reflect the tripeptide repeat structures that occur at higher frequencies in alleles from either continent (section 2.3.9).

Antigen	Pop ⁿ	All alleles	K1-like	MAD20 -like	RO-33-like	MR-like
PVHA _g	All	0.85 (0.47-1)	0.96 (0.65-1)	0.84 (0.59-1)	0.78 (0.54-0.78)	0.58 (0.47-0.65)
PVHA _{af}	All	0.79 (0.33-1)	1.0 (0.74-1)	0.71 (0.33-1)	0.96 (0.54-1)	0.72 (0.65-0.75)
PVHA _{as}	All	0.78 (0.47-1)	0.82 (0.48-1)	0.97 (0.62-1)	0.78 (0.54-1)	0.58 (0.47-0.63)
PVHA _f	All	0.93 (0.33-1)	0.98 (0.78-1)	0.63 (0.33-1)	0.96 (0.54-1)	0.51 (0.51-0.63)
PVHA _{af}	Africa	1.0 (0.33-1)	1.0 (0.74-1)	0.85 (0.33-1)	1.0 (0.54-1)	0.72 (0.65-0.75)
PVHA _f	Africa	0.98 (0.33-1)	0.98 (0.78-1)	0.79 (0.33-1)	1.0 (0.54-1)	0.51 (0.51-0.63)
PVHA _{as}	Asia	1.0 (0.54-1)	1.0 (0.75-1)	0.97 (0.68-1)	1.0 (0.54-1)	0.58 (0.58-0.58)
PVHA _f	Asia	0.78 (0.33-0.78)	0.98 (0.89-1)	0.54 (0.33-1)	0.78 (0.78-1)	0.51 (0.51-0.51)

Table 3.2 Comparison of coverage of MSP-1 block 2 sequences by novel designs for polyvalent hybrid antigens. Nonamer amino acid sequences from translated reads were used to create polyvalent hybrid antigen sequences using an algorithm that combines these nonamer sequences based on their frequency in the population. One antigen (global polyvalent hybrid antigen, PVHA_g) was designed using nonamers from all samples in the Pf3k data and two antigens were designed using nonamer sequences from African and Asian Pf3k samples (African polyvalent hybrid antigen, PVHA_{af} and Asian polyvalent hybrid antigen global, PVHA_{as}, respectively). These antigens were then tested against peptide sequences in the tSRA dataset (section 2.3.9). The number of unique nonamers from each tSRA sequence which occurred in the tested antigen sequence was determined as a fraction of all unique nonamer sequences present in each tSRA sequence. The median fraction of nonamers present in the antigen sequences are shown, with the range in parenthesis for each antigen tested against *de novo* assembled sequences from all regions for all allelic types and each allelic type individually. The region specific antigens were also tested against *de novo* assembled sequences from the region for which the antigen was designed. The manually designed polyvalent hybrid antigen f (PVHA_f) (Tetteh and Conway, 2011) is included for comparison.

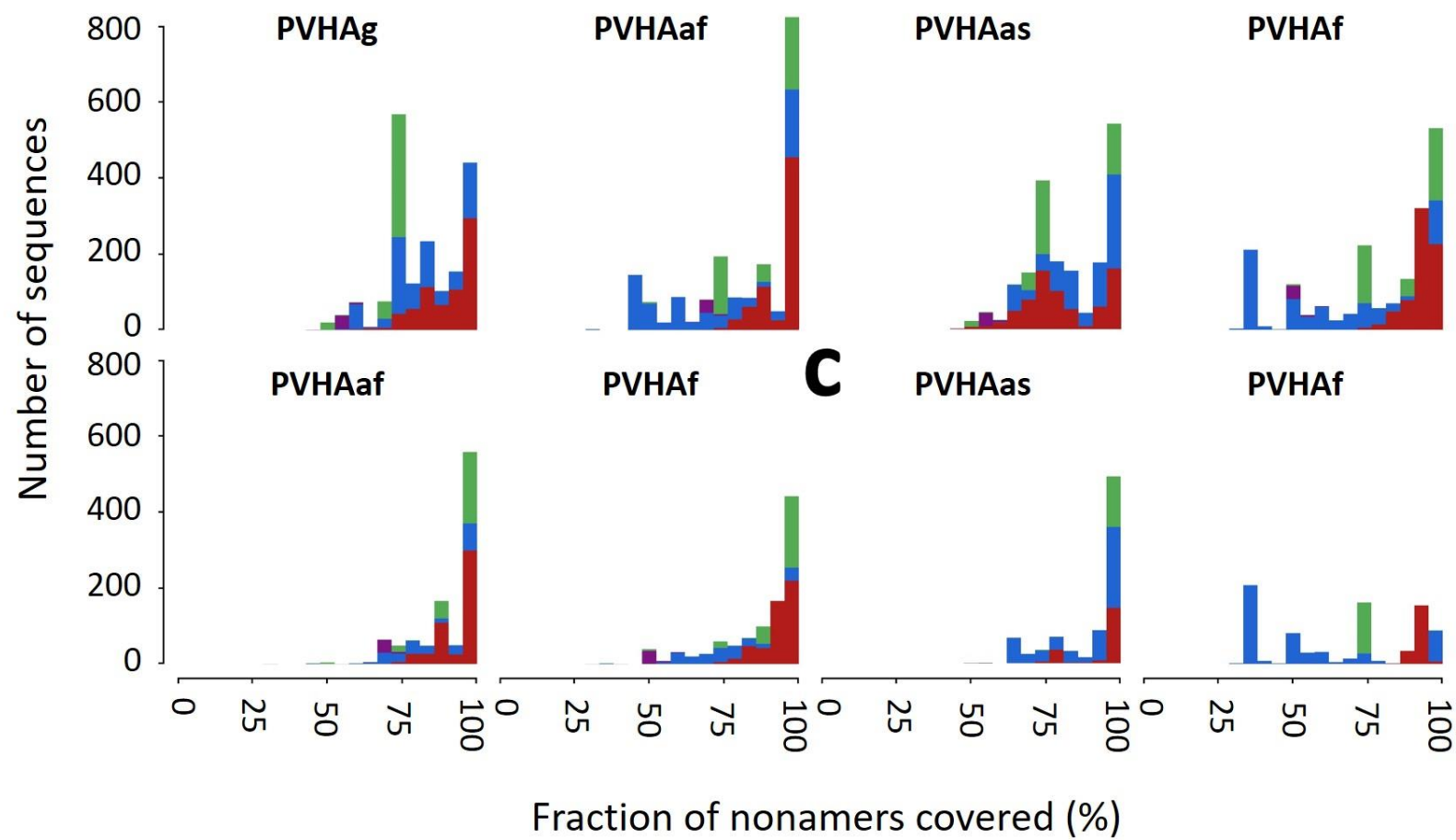
a**b**

Figure 3.7 Nonamer peptide coverage of allele sequences in the tSRA dataset by proposed polyvalent hybrid antigens. Short reads from Pf3k samples were aligned to the *msp1b2RefLib* (appendix 7.4) and translated (section 3.2.2). Nonamer sequences from these translated reads were used to design polyvalent hybrid antigens (section 3.3.3). Polyvalent hybrid antigens were designed from all nonamer sequences (global polyvalent hybrid antigen, PVHAg) and from African (African polyvalent hybrid antigen, PVHAaf) and Asian nonamer sequences (Asian polyvalent hybrid antigen, PVHAas). These antigens were tested *in silico* against each *de novo* assembled sequence in the tSRA dataset (section 2.3.9) by determining the fraction of nonamer sequences represented in the antigen. Histograms show the frequency distribution of the percentage of nonamers from each tSRA sequence that are present in each antigen. Bars are coloured by sequence type: K1-like (red), MAD20-like (blue), RO-33-like (green) and MR-like (purple). Polyvalent hybrid antigen F (PVHAF), which has a K1-super repeat designed manually, combined with Wellcome (MAD20-like) and RO-33 sequences, was tested against tSRA sequences in the same manner for each population. (A) All antigens were compared against all sequences in the tSRA (n=1522). (B) The African antigen design (PVHAaf) was also compared against African tSRA sequences (n=787) and, (C), the Asian antigen (PVHAas) was compared against Asian sequences (n=735).

3.4 Discussion

Short read data generated from LRD sequences (section 2.3.1) was used to validate the translation of short reads based on alignment to the sequence library described previously (section 2.3.4). No spurious amino acid sequences resulted from the translation of reads using the frame suggested by their alignment, demonstrating the validity of this approach. Furthermore, over 80% of unique nonamer sequences present in the LRD dataset were recovered by this approach, indicating its potential for extracting nonamer amino acid sequence data from short reads containing polymorphic repeat sequences. The nonamer sequences that were not detected from this approach originated from rare sequences that do not align to the *msh1b2RefLib*, showing a key weakness of this approach. However, the design of hybrid vaccine antigens will be based on the most common sequence variants and this method can therefore be applied here.

Applied to the short read data from the Pf3k project, translation of the reads resulted in a 10-fold larger number of unique nonamer sequences than those present in the LRD sequences. This is in part due to the larger number of sequences present in this data set which contains almost three times as many samples and, due to the presence of mixed genotype infections, at least four times as many sequences (section 2.3.5). However, the increase in the unique number of nonamer sequences is doubtless also driven by translation of reads containing sequencing errors that would not be present in the *in silico* generated reads. These reads were not excluded from the analysis, as subsequent steps to analyse nonamer sequences ignored nonamers present at low frequencies. The analysis of regional nonamer frequencies (figure 3.1) reflected the pattern seen in the *de novo* assembled sequences (section 2.3.9) and in previous studies employing long read sequencing (Tanabe, 2013, Noranate et al., 2009), which suggests the method is valid. The fact that the most frequent K1-like tripeptide repeat nonamers seen in Pf3k samples from Africa are represented by the 12-mer peptides that were recognised by antibodies from multiple sera samples collected from adults living

in the same region (Tetteh et al., 2005b) is further evidence that the approach of translating short read sequence to determine nonamer frequency is appropriate.

The design of polyvalent hybrid antigen proteins aims to incorporate the maximum amount of sequence diversity into the shortest possible antigen sequence in order to give an antigen that can be easily produced and has the potential to elicit an antibody response against a wide range of alleles. To do this, an algorithm was developed that ranks nonamer sequences based on their frequency in the population and their degree of overlap with other nonamer sequences. The highest ranked nonamers are then incorporated into a polyvalent hybrid antigen construct as they offer the greatest increase in representation of sequence diversity with the minimal increase of construct length. The algorithm uses a linear scaling factor, termed *sfp*, which adjusts the weighting of nonamer frequency and overlap in determining the score for each nonamer (equation 3.1). This parameter was optimised to produce antigen sequences containing the highest number of nonamer sequences for each LRD sequence. The antigen produced by the optimised algorithm compared favourably with the previously described PVHAf (Tetteh and Conway, 2011) when analysed against LRD sequences.

The established difference in the frequencies of the allelic types (section 2.3.6), along with the greater diversity present in the African sequences, leads to a marked difference in nonamer frequencies between the two continents (figure 3.1). With this in mind, the algorithm was used to design three separate antigens; one global antigen (PVHAg) from all the data and two regional antigens using data from African isolates (PVHAaf) or Asian isolates (PVHAas) only. As expected on the basis of the allele frequencies, the African and Asia antigens contained greater proportions of K1- and MAD20-like sequences respectively.

All three antigens designed by an algorithm to include nonamer sequences based on their frequency and the increase in antigen length required to include them contain a higher proportion of MAD20-like allele sequences than PVHAf (figure 3.7, table 3.2), which is to be expected as this antigen

contains only a single MAD20-like allele (Tetteh and Conway, 2011). The MAD20-like hybrid sequences designed by this algorithm could replace the Wellcome (MAD20-like) allele that is in the present antigen. This update to the PVHaf would be likely to be of greater importance if the antigen was to be tested as a vaccine in Asia, where the increased complexity of MAD20-like repeat structures (section 2.3.9) may enable vaccine escape if only single MAD20 allele is included in a vaccine formulation. It is unfortunate that the sequence of another proposed MSP-1 block 2 hybrid antigen, which contains a synthetic MAD20 sequence designed to present a wider range of MAD20-like epitopes has not been made publicly available and so could not be compared to the antigen sequences designed algorithmically (Cowan et al., 2011).

PVHaf contains a synthetic K1-like repeat that is designed to incorporate a range of naturally occurring variants; encouragingly, the antigens designed by the algorithm are predicted to cover a comparable degree of K1-like sequences. This demonstrates the successful use of short read sequence data to design antigens that incorporate a range of sequence diversity without the need to first assemble sequences. Given that the reference sequence library contains a modest number (15) of different allele sequences, it is therefore feasible that limited long read sequencing of complex repeat regions would allow the capture of population-wide nonamer frequencies from short read data and the subsequent design of hybrid antigens that represent the naturally occurring sequence diversity. Given the prevalence of polymorphic repetitive sequences in *P. falciparum* vaccine candidates predicted to present linear B-cell epitopes (Feng et al., 2006, Guy et al., 2015), this tool has the potential to aid in the design of vaccine candidates that condense a large number of allelic variants into a relatively short protein sequence.

Antigens comprising the two main variants of another merozoite surface protein, MSP-2, have already been proposed and shown to elicit cross-strain immune responses (Krishnarajuna et al., 2016). The algorithm described here could be used to incorporate greater sequence diversity into a hybrid MSP-2 antigen and could also be applied to other *P. falciparum* antigens predicted to present

polymorphic, linear B-cell epitopes, such as Merozoite Surface Protein Duffy Binding Like 1 (MSPDBL-1), MSPDBL-2 and Serine Repeat Antigen 5 (SERA-5).

Chapter 4 - Experimental approaches toward producing recombinant monoclonal antibodies against MSP-1

4.1 Introduction

Multiple studies have shown a correlation between presence of IgG against MSP-1 block 2 and protection from malaria (reviewed in (Fowkes et al., 2010)), but a direct causal effect and the mode of action of these antibodies is still unclear. Whilst studies control for exposure to malaria, it is still possible that IgG against MSP-1 block 2 are markers of exposure and do not directly convey protection. Vaccination will only boost antibodies against the antigen or antigens present in the vaccine and it is therefore important to establish that naturally-acquired antibodies against a specific antigen are efficacious prior to development as a vaccine target.

Immunization of *Aotus* monkeys with MAD20-like MSP-1 block 2 resulted in control of parasitaemia for two out of four individuals infected with a homologous parasite strain, indicating that antibodies against MSP-1 block 2 can provide protection in this model of infection (Cavanagh et al., 2014).

However, anti MSP-1 block 2 antibodies obtained both by animal immunization and affinity purification of human sera did not show significant, direct inhibition of *in vitro* parasite growth (Cowan et al., 2011, Galamo et al., 2009), implying that antibodies against this region of MSP-1 have an indirect mode of action. Indeed, antibodies affinity purified from clinically immune adult sera from Côte d'Ivoire against a K1-like MSP-1 block 2 antigen showed allele-specific inhibition of parasite growth in the presence of naïve monocytes, demonstrating the potential for anti-MSP-1 block 2 antibodies to trigger cellular inhibition of parasites (Galamo et al., 2009). Rabbit antibodies raised by immunisation with MAD20-like MSP-1 block 2 antigens were shown to inhibit invasion of merozoites expressing homologous MSP-1 block 2 in the presence of active human complement proteins, suggesting that activation of the complement cascade by anti-MSP-1 block 2 antibodies results in either blocking of invasion, lysis of the merozoite or both (Boyle et al., 2015). Whilst these studies give some evidence for the efficacy of anti-MSP-1 block 2 IgG, antibodies against all three

allelic types have not been tested for their efficacy against parasites bearing the homologous MSP-1 block 2 allelic type. Furthermore, only the study looking at antibody dependent cellular inhibition of parasite growth used human antibodies. It is therefore necessary to further investigate the efficacy of human antibodies against all allelic types of MSP-1 block 2 to evaluate this promising antigen as a vaccine target.

Specific human antibodies can be purified from sera collected from individuals with clinical immunity by means of their affinity for a specific antigen (Wofsy and Burr, 1969). However, this technique has a number of drawbacks: the antibodies with the highest affinity will never be recovered as they will remain bound to the antigen; poor yields can result in the need for large volumes of sera, often only achievable by pooling sera from multiple donors, potentially obscuring variation between individuals; the resulting antibodies are a polyclonal mix that cannot be replicated; and it can never be certain that low levels of non-specific antibodies remain in the preparation. For these reasons, production of human monoclonal antibodies was attempted.

Human monoclonals can be produced by immortalisation of B-cells collected from exposed donors (Traggiai, 2012). Screening of antibodies expressed by the resultant cultures can then determine reactivity to antigens of interest. This approach has been used for the production of monoclonals recognising *P. falciparum* merozoite surface proteins (MSPs) (Maskus et al., 2015, Maskus et al., 2016, Stubbs et al., 2011) and other blood-stage antigen targets (Sirima et al., 2016, Barfod et al., 2007, Berzins et al., 1985, Udomsangpetch et al., 1986). Reverse transcription of B-cell mRNA and cloning of sequences encoding FAb fragments into phage display libraries represents another approach for producing human monoclonals and has been applied to blood stage antigens including MSP-1 block 2 (Lundquist et al., 2006, Cheng et al., 2007, Sowa et al., 2001). Sorting of antigen-specific cells has the advantage of avoiding labour intensive cell culture and screening and has the potential to produce large numbers of monoclonals (Muellenbeck et al., 2013). This enables screening of a greater number of memory B-cells, and was therefore selected for use in this project

due to the predicted low frequency of MSP-1-specific memory B-cells in malaria exposed adults (Nogaro et al., 2011).

Human B-cells recognise antigen epitopes via cell-surface receptors (B-cell receptors, BCRs), consisting of two heavy and two light immunoglobulin chains. During B-cell development, the immunoglobulin heavy (IgH) chain variable region is formed by the recombination of one of over 51 variable region (V) with one of 25 diversity (D) and one of 6 joining (J) genes present at the IgH locus on chromosome 14 (Watson et al., 2013, Murphy et al., 2008, Davis et al., 1980) figure 4.1).

Recombination of one of 30-40 V genes with one of 4-5 J genes present at either the Ig κ (chromosome 2) or Ig λ (chromosome 22) loci occurs to produce a light chain variable region (figure 4.1). Huge diversity is created by the possible combinations of these genes and by the addition of non-templated nucleotides at the joining sites. The variable regions of the heavy and light immunoglobulin chain are responsible for binding antigen, so it is this vast diversity created through V(D)J recombination that creates unique naïve B-cells and forms the basis of their recognition of a vast number of antigens. Following V(D)J recombination, mRNA transcripts from the IgH and either the Ig κ or Ig λ loci are expressed and spliced such that the V(D)J genes encoding the variable region are joined to the exons encoding the constant domains (figure 4.1).

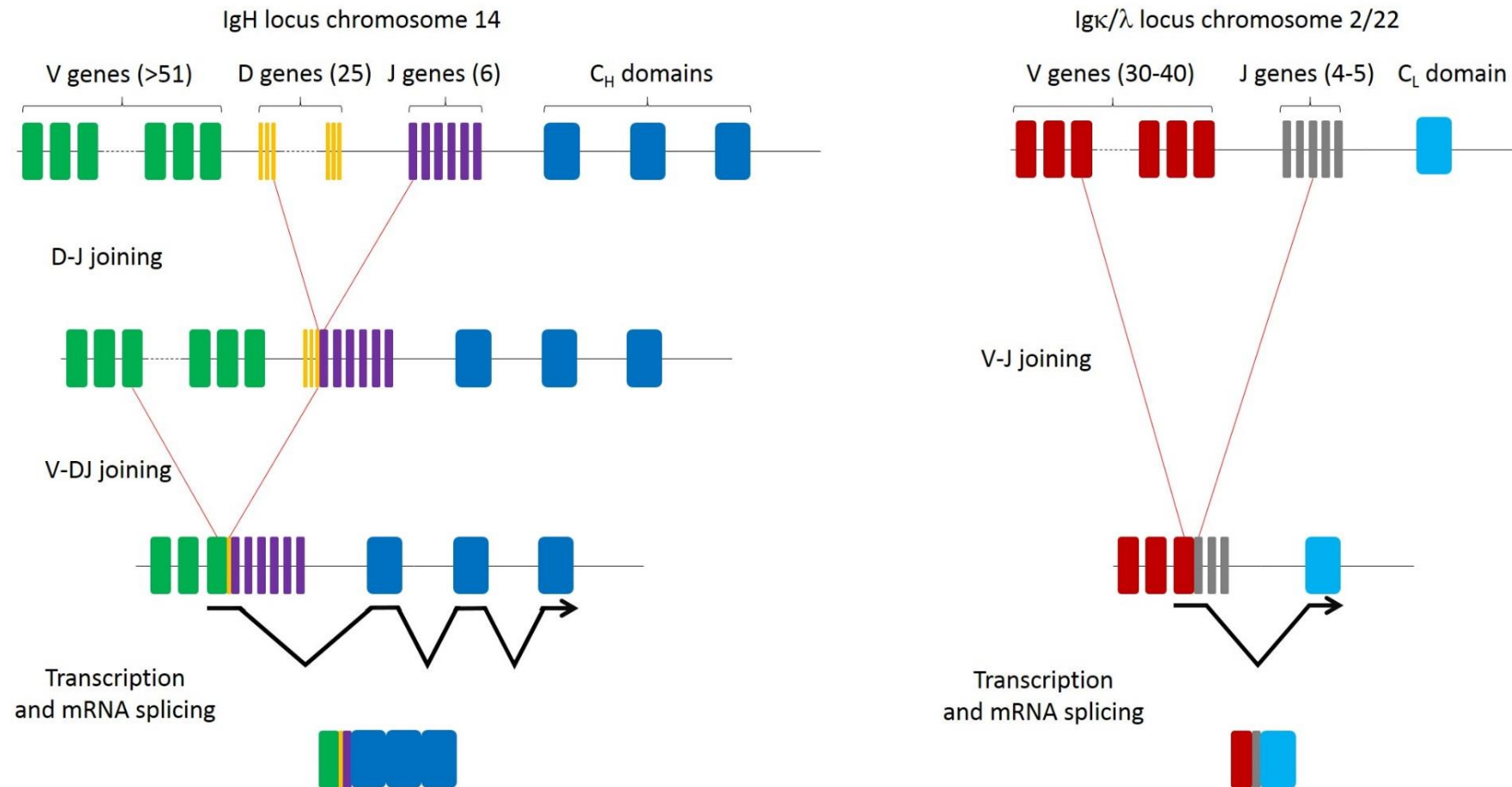


Figure 4.1. Schematic representation of recombination at the IgH and Igκ/λ loci to produce heavy and light chain immunoglobulin transcripts. During development of the pre B-cell in the bone marrow, the IgH locus on chromosome 14 undergoes two steps of recombination to join a single D gene (yellow) with a single J gene (purple) and then to join the recombined DJ genes with a single V gene (green). Maturation of the pre B-cell to a pro B-cell is marked by the recombination of either the Igκ or Igλ locus, in which a single V gene (red) is joined to a single J gene (grey). Splicing of mRNA transcripts from both heavy and light chain loci results in a transcript encoding the unique variable region followed by three (heavy chain) or a single (light chain) constant region. Figure adapted from (Janeway et al., 2001).

In frame recombination results in the expression of the light and heavy immunoglobulin chains consisting of a variable domain, encoded by recombined V(D)J genes and one (for light chains) or three (for heavy chains) constant domains; two heavy and two light chains are joined by disulphide bonds to form a BCR (figure 4.2). Within both heavy and light chain variable domains there are three complementarity determining regions (CDRs) which contain the majority of residues that contact the antigen separated by four framework regions (FR) that contain fewer antigen binding residues and are involved in forming the immunoglobulin domain structure (Davies et al., 1990) figure 4.2). It is therefore the CDRs that are primarily involved in antigen binding.

If the BCR binds to foreign antigen, the naïve B-cell is activated and undergoes clonal expansion. Some of the cells resulting from this expansion will become plasma cells that secrete the BCR as soluble immunoglobulin, other cells will undergo affinity maturation in the germinal centre, a process by which additional diversity is generated by targeted mutation of the recombined V(D)J genes, known as somatic hypermutation (SHM), and selective retention of B-cells with higher affinity for the specific antigen (Jacob et al., 1991, Pham et al., 2003). This affinity-matured population will go on to form memory B-cells that persist in the absence of antigen and form the basis of humoral immune memory (McHeyzer-Williams and McHeyzer-Williams, 2005).

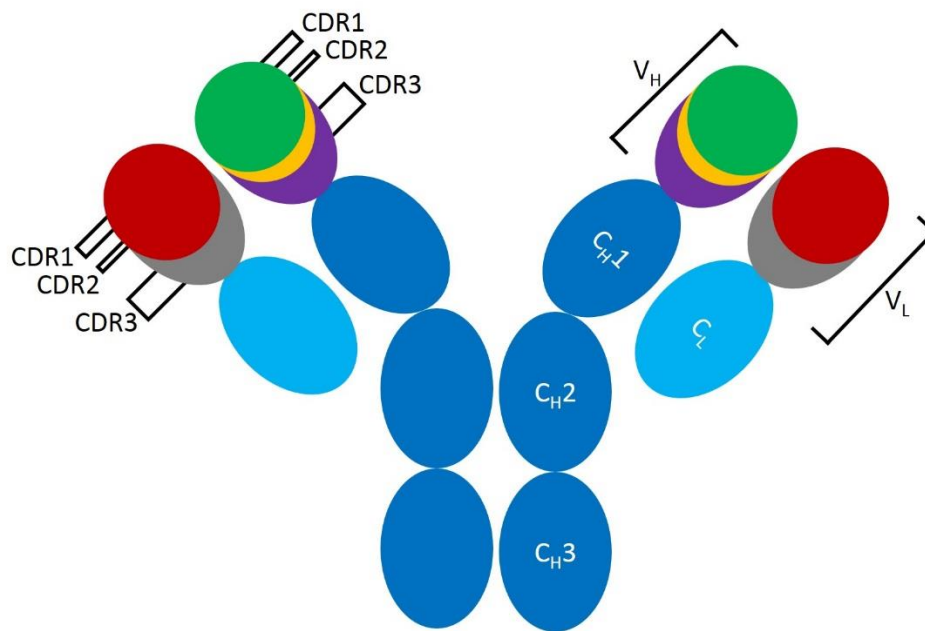


Figure 4.2 Schematic representation of BCR structure. The BCR consists of two heavy and two light chains comprising four and two immunoglobulin domains respectively. Each chain has an N-terminal variable domain encoded by joining of a V (green) D (yellow) and J (purple) gene for the heavy chain variable domain (V_H) and by joining of a V (red) and J (grey) gene for the light chain variable domain (V_L) (Schroeder and Cavacini, 2010). Within the variable domains there are three complimentary determining regions (CDR1-3) which are primarily involved in antigen recognition (Davies et al., 1990). Figure adapted from (Janeway et al., 2001).

The presence of MSP-1 specific memory B-cells, as detected by ELISpot assay, has been demonstrated in children living in a malaria endemic region of Kenya, and these cells persist longer than detectable antibody responses for children living in a region of Kenya (Ngerenya) in which a drop in malaria transmission meant that they were previously exposed to malaria but had not recently been exposed (Ndungu et al., 2012). Although the titres of antibodies against *P. falciparum* merozoite antigens do not always correlate with frequencies of memory B-cells recognising that antigen (Nogaro et al., 2011, Ndungu et al., 2012), the fact that MSP-1 block 2 antibodies can be detected in a high proportion (47%) of adults, who have a history of malaria, but no concurrent infection, shows that MSP-1 block 2 elicits an immune response in exposed adults (Polley et al., 2003b). Assuming a normal B-cell response to this antigen in adults, this indicates that MSP-1 block 2-specific memory B-cells will be present in adults resident in malaria endemic settings, such as Kintampo, Ghana, the site for this study (Owusu-Agyei et al., 2012).

Fluorescence assisted cell sorting (FACS) can be used to separate cells based on their fluorescent properties, and thus to isolate fluorescently labelled cells. The use of fluorescent antigen to purify antigen-specific B-cells by this technique is well-established (Greenstein et al., 1980, Hayakawa et al., 1987, Julius et al., 1972, McHeyzer-Williams et al., 2000, Townsend et al., 2001). This method has been used to isolate memory B-cells recognising a malaria vaccine candidate through chemical linkage of a fluorophore to the recombinant GLURP-MSP-3 hybrid antigen, GMZ2 (Muellenbeck et al., 2013). Work with tetanus toxin has shown that the tetramerisation of antigens, via biotinylation and subsequent binding to fluorescently labelled streptavidin, can increase the specificity of B-cell labelling and thus increase the proportion of B-cells isolated that encode antigen-specific immunoglobulin (Franz et al., 2011).

BCR-encoding mRNAs from single, isolated B-cells can be reverse transcribed and the V(D)J genes encoding the variable region can be amplified by nested PCR (Tiller et al., 2008). The amplified fragments can be sequenced and cloned into expression vectors for recombinant expression of

antigen-specific monoclonal antibodies in a human cell line (Dodev et al., 2014). This approach can therefore be used to reproduce high-affinity, naturally occurring, human MSP-1 block 2 specific antibodies.

In order to produce recombinant human monoclonal antibodies against MSP-1 block 2, biotinylation and tetramerisation of a hybrid antigen representing all MSP-1 block 2 types (polyvalent hybrid (PVH) antigen F (Tetteh and Conway, 2011)) was attempted. The approach used to produce these tetramers was unsuccessful due to a failure to chemically biotinylate the antigen. In order to assess whether *P. falciparum* antigen tetramers could be used to isolate specific memory B-cells from exposed individuals, a full length MSP-1 antigen was used that had already been biochemically biotinylated. Once isolated, the variable regions of the heavy and light chain immunoglobulin expressed by individual MSP-1-specific B-cells were analysed. These antibodies could then be used to map the MSP-1 block 2 epitopes that are recognised by B-cells and also be used in *in vitro* assays to investigate the efficacy of antibodies recognising this antigen.

4.2 Materials and methods

4.2.1 Polyvalent hybrid (PVH) MSP-1 antigens

Microbeads with *E. coli* BL21 cells transfected with pET-15b-PVH antigen F plasmids (Tetteh and Conway, 2011) (a kind gift from Dr Kevin Tetteh) were used to inoculate starter cultures of 5 mL of lysogeny broth (LB, (Bertani, 1951)) with 100 µg mL⁻¹ ampicillin and incubated overnight with at 37°C. 6.25 µg of pET-15b-PVH antigen F plasmid was purified from one 5 mL culture using NucleoSpin Plasmid purification kit (Macherey-Nagel). Purified plasmid was sequenced using the BigDye Terminator v3.1 Cycle Sequencing kit (Applied Biosystems) using manufacturer's instructions. The pET forward sequencing primer (5'-TAATACGACTCACTATAGGG-3', 10 µM) was added to the sequencing reaction, performed under the following conditions: 96 °C for 1 minute followed by 25

cycles of 96 °C for 30 seconds, 50 °C for 5 second and 60 °C for 3 minutes and 45 seconds. Following sodium acetate/ethanol precipitation to remove unincorporated dyes, reaction products were dissolved in highly de-ionised formamide buffer (Applied Biosystems) and analysed by electrophoresis using an ABI3730 sequencer. All DNA sequence chromatograms were examined and readings were corrected by eye where necessary using Finch TV (Geospiza).

Purified pET-15b-PVH antigen F plasmid was mutated by polymerase incomplete primer extension (PIPE) (Klock and Lesley, 2009) using PVHAntigenF-S26C-3pOLfwd (sequence: 5'–GACCCATGAATGCTATCAGGAACTGGTTAAAAAAGTGAAG–3') and PVHAntigenF-S26C-3pOLrev (sequence: 5'–TCCTGATAGCATTCATGGGTCACGGATCCGGTA–3') primers designed to introduce an A → T mutation at position 78 of the PVH antigen F sequence and to linearise the pET-15b-PVH antigen F plasmid with complementary 3' overhangs. 25 ng of plasmid DNA was combined with forward and reverse primers (0.5 µM) and Phusion Flash High-Fidelity PCR Master Mix (Thermo Scientific), containing Phusion Flash II DNA Polymerase and exposed to an initial denaturation step of 98 °C for 10 seconds followed by 30 cycles of denaturation at 98 °C for 1 second, followed by annealing and extension at 72 °C for 90 seconds. Following thermocycling, parental plasmid was digested with *DpnI* in CutSmart buffer (New England Biolabs) for 15 minutes at 37 °C with mixing. Mutant plasmids were used to transfect *E. coli* NEB 10 β cells (New England Biolabs) via heat-shock and transfectants were selected by overnight growth on ampicillin plates. 5 mL cultures of LB with 100 µg mL⁻¹ ampicillin were inoculated with 9 individual colonies and grown overnight at 37 °C. Plasmids were purified and sequenced as above; corrected sequences were aligned using Clustal Omega (Sievers et al., 2011). Purified plasmid bearing the desired mutation was used to transfect *E. coli* BL21 cells made chemically competent by use of Mix n Go kit (Zymoresearch) according to manufacturer's instructions. Successfully transfected cells were selected by growth on ampicillin plates and one colony was used to inoculate one 5 mL starter culture of LB with 100 µg mL⁻¹ ampicillin; a second start culture was inoculated with a microbead containing *E. coli* BL21 cells transfected with parental pET-15b-PVH antigen F plasmids. After overnight growth at 37 °C 4 mL of starter cultures was used

to inoculate two 1 L cultures of ZYM-5052 autoinduction media (Studier, 2005) complemented with $100\text{ }\mu\text{g mL}^{-1}$ ampicillin. Following growth overnight at $37\text{ }^{\circ}\text{C}$ cells were pelleted, freeze-thawed and then mechanically lysed using a FastPrep™ with Matrix beads B (MP Biomedical). Cell lysates were clarified by centrifugation at $15,000\text{ g}$ and diluted 1:1 in PBS with 10 mM imidazole. His-tagged protein was bound to Ni-Nitrilotriacetic acid beads (Thermo Scientific). Following washing in 15 mL of PBS with 25 mM imidazole, PVH antigens were eluted in 3 mL of 250 mM imidazole in PBS. Eluted protein was transferred into PBS and concentrated using spin columns (molecular weight (M_w) cut-off 10 kDa , Amicon). Protein concentration was estimated by reaction of $5\text{ }\mu\text{L}$ samples with 5 mL of Protein Assay Dye Reagent (Bio-Rad) for 10 minutes at room temperature prior to measurement of absorbance at 595 nm . Estimation of protein concentration by comparison to a standard curve constructed with bovine serum albumin (BSA, Sigma) showed that 1.45 mg of mutated PVH_{S26C} antigen F and 2.01 mg of PVH antigen F had been produced. Samples of PVH antigens were denatured and analysed on 10% sodium dodecyl sulphate-polyacrylamide gel electrophoresis (SDS-PAGE, figure 4.6). PVH antigen F_{S26C} was reacted with a 20-fold excess of maleimide-(polyethylene glycol (PEG))₁₁-biotin (Thermo Scientific) for 2 hours at room temperature. Excess maleimide-PEG₁₁-biotin was removed using spin columns (M_w cut-off 10 kDa , Amicon).

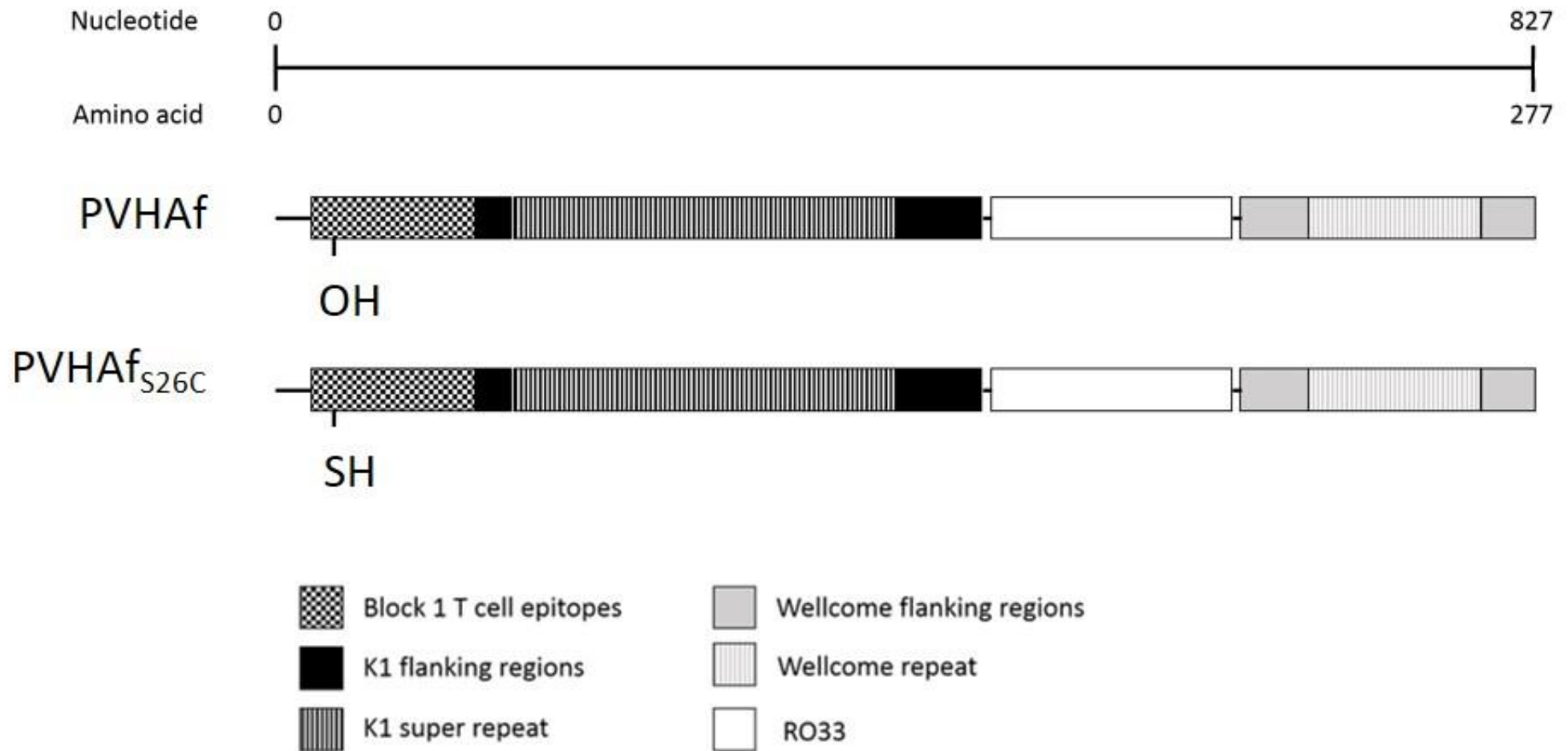


Figure 4.3 Schematic representation of the introduction of a free sulfhydryl group into the polyvalent hybrid (PVH) MSP-1 antigen. In order to introduce a free sulfhydryl (SH) group that could be crosslinked to a biotin moiety by formation of a thioether with a maleimide group, the polyvalent hybrid antigen F (PVHAf), encoding MSP-1 block 1 T-cell epitopes (black and white chequered) followed by the synthetic K1 super repeat (Tetteh et al., 2005a) (black and white striped) flanked by the K1 non-repeat regions (black) linked to the RO-33 allelic sequence followed by the MAD20-like Wellcome allelic sequence, was mutated to introduce substitute a cysteine residue at position 26 in the first block 1 T-cell epitope in place of the original serine. Figure adapted from (Tetteh and Conway, 2011).

4.2.2 Full length MSP-1 antigen

Supernatant from HEK293E cells transiently co-expressing BirA biotinylation enzyme and a full-length, his-tagged *P. falciparum* 3D7 MSP-1-rat CD4 fusion protein with a C-terminal BirA biotinylation site (MSP-1 biotinylation site linker histidine tag (MSP1-BLH) (Crosnier et al., 2013)), was a kind gift from Dr Gavin Wright. Imidazole and NaCl were added to supernatants to a final concentration of 10 mM and 100 mM respectively. Supernatants were then bound to a HisTrap HP 1 mL column (GE Healthcare Life Sciences) overnight. Column was washed in phosphate buffer (16.4 mM K_2HPO_4 , 3.96 mM KH_2PO_4 , 20 mM imidazole, 100 mM NaCl, pH 7.4) for 5 column volumes following return of A_{280} to baseline. His tagged proteins were eluted in 100 mM imidazole in phosphate buffer (16.4 mM K_2HPO_4 , 3.96 mM KH_2PO_4 , 100 mM imidazole, 100 mM NaCl, pH 7.4). Eluted fractions under the A_{280} peak were collected and 100 μ L samples of eluted fractions were reacted with 5 mL of Protein Assay Dye Reagent (Bio-Rad) for 5 minutes at room temperature prior to measurement of absorbance at 595 nm; protein concentration was estimated by comparison to a standard curve constructed with bovine serum albumin (BSA, Sigma).

Samples of eluted fractions containing over 1 μ g mL⁻¹ protein were analysed by SDS-PAGE (figure 4.4a). Protein bands, from a simultaneously run, unstained SDS-PAGE gel were transferred to a nitrocellulose membrane (Amersham), incubated for 30 minutes at room temperature in blocking buffer (3% milk powder, 0.1% TWEEN-20 in 25 mM Tris base, 200 mM NaCl) before probing with rabbit α his tag antibody (1:1000 dilution in blocking buffer, Raybiotech), washing in TBS/TWEEN (0.1% TWEEN-20 in 25 mM Tris base, 200 mM NaCl) and detection with DyLight™ 680 labelled goat α rabbit IgG antibody (100 ngmL⁻¹, KPL). Unbound labelled antibody was removed by washing in TBS/TWEEN and read fluorescence was read using an Odyssey® imager (LI-COR Biosciences) at 700 nm (figure 4.4b). Eluted fractions with over 0.25 mg mL⁻¹ protein were pooled and dialysed against PBS. Post dialysis protein yield was estimated by Bradford assay (see above) to be 1.4 mg purified Bio-MSP1-BLH

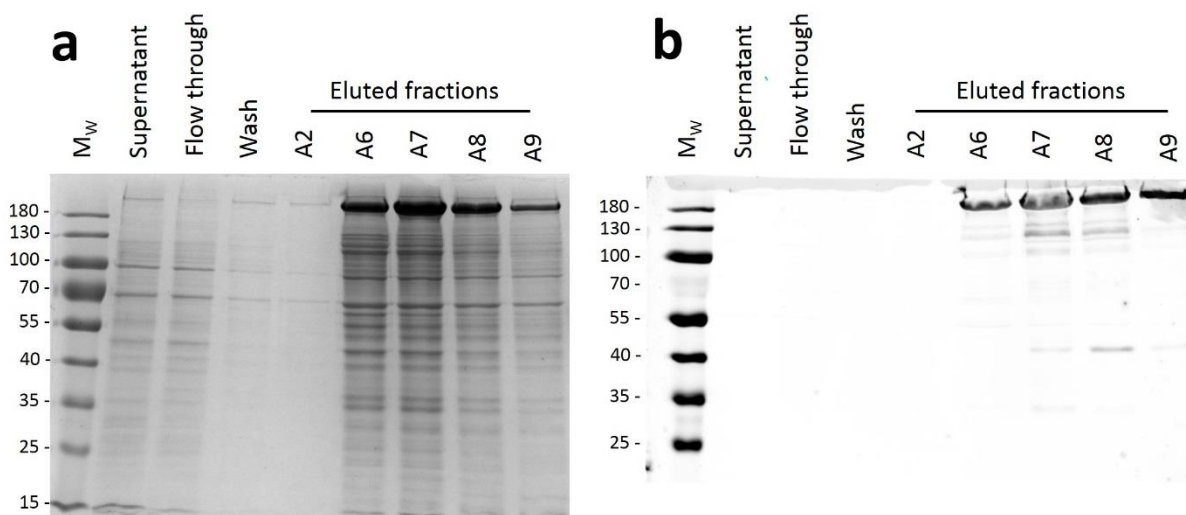


Figure 4.4 Analysis of purified Bio-MSP1-BLH. (A) Coomassie stained gel showing purification of Bio-MSP1-BLH and (B) Western blot showing presence of his-tagged Bio-MSP1-BLH protein in eluted fractions. 5 μ L PageRuler™ Prestained Ladder (lane 1) 10 μ L of pre-purification supernatant (lane 2), 10 μ L flow through (lane 3), 10 μ L wash (lane 4), 10 μ L of fraction A2 (lane 5) and 10 μ L each of fractions A6-A9 (lanes 6-9) were loaded onto a 10% SDS-PAGE gel and run at 200V for 40 minutes. Gels were either washed in H₂O, stained with Bio-Safe™ Coomassie stain (Bio-rad) and destained in H₂O (a) or transferred onto a nitrocellulose membrane (Amersham), probed with rabbit α his tag antibody (1: 10 000 dilution Raybiotech) and detected with DyLight™ 680 labelled goat α rabbit IgG antibody (100 ngmL⁻¹, KPL) (b). Blot was washed and read on an Odyssey® imager (LI-COR Biosciences) at 700 nm.

4.2.3 Tetramerisation of antigens

For optimisation of antigen tetramerisation, purified antigens were mixed with streptavidin (SA, Sigma) at a range of molar ratios and incubated at 4 °C for 30 minutes. Samples were loaded onto a 4-12% NuPAGE gel (Novex) and run for 1 hour and 30 minutes. Gels were stained washed three times for 5 minutes in H₂O, stained with Bio-Safe™ Coomassie stain (Bio-rad) for 1 hour and destained in H₂O for 30 minutes.

For preparation of MSP-1 antigen tetramers for use in labelling B-cells (section 4.2.5), purified BioMSP1-BLH was mixed with streptavidin-R-phycoerythrin (SAPE, Molecular Probes) in an 8:1 molar ratio and incubated on at 4 °C for 30 minutes. Aggregates were removed by centrifugation prior to dilution to a final concentration of 0.125 μ g mL⁻¹ in PBS.

4.2.4 Sample collection

Healthy adult males aged 18 – 49 years (median age 31 years) who had lived in Kintampo North Municipal district for over 4 months were recruited (a description of the ecology and malaria transmission of Kintampo can be found (section 2.2.6) above). Adult donors on the London School of Hygiene & Tropical Medicine anonymous blood donors' register, who had no history of malaria were also recruited. 60 mL of venous blood was collected and samples sent for haematology. Plasma was separated by centrifugation and tested for reactivity to five MSP-1 block 2 antigens (section 2.2.8) by ELISA as described above (section 2.2.9). Peripheral blood mononuclear cells (PBMC) were separated by density gradient centrifugation with Lymphoprep™ (Alere Technologies) and cell count and viability by trypan blue exclusion was assessed prior to cryopreservation at -80°C in 10% dimethyl sulfoxide (DMSO) in heat-inactivated foetal bovine serum (FBS). Ethical approval for this study was granted by the London School of Hygiene & Tropical Medicine (reference 9161), Kintampo Health Research Centre (reference 2015-4) and the Ministry of Health, Ghana.

4.2.5 Preparation of B-cells

Cell counts, viability and the percentage of CD19⁺ were used to rank samples by the predicted number of B-cells. For exposed (Kintampo) samples only those for which plasma reacted to one or more MSP-1 block 2 samples were thawed. B-cells were enriched through magnetic depletion of non-B-lymphocytes and erythrocytes using EasySep™ human B-cell enrichment kit (STEMCell). Enriched B-cells were stained with FVS-780 (BD Bioscience) prior to blocking of F_C receptors with TruStain™ (BioLengend). Cells were then labelled with anti-CD19-BB515, anti-CD27-APC, anti-CD3-PerCP, anti-CD14-PerCP (BD Bioscience) and anti-CD16-PerCP (BioLegend) at recommended concentrations before binding of MSP-1-SAPE antigen tetramers at 4°C for 30 minutes. Unbound antibodies and antigen tetramers were removed by washing prior to loading of cells onto a BD FACS Aria II (BD Biosciences). Individual, CD19⁺, CD27^{high}, CD3/CD14/CD16⁻, MSP-1-SAPE⁺ cells were sorted into a 96 well PCR plate containing 4 µL of lysis buffer (0.5X PBS, 10 mM dithiothreitol (DTT)

and 2U μL^{-1} RNAsin[®] (Promega) per well. Immediately following sorting, plates were sealed and stored at -80°C.

4.2.6 Amplification of Ig gene variable regions

The PCR strategy for amplification of Ig gene variable regions from B-cell messenger ribonucleic acid (mRNA) published by Wardemann and Kofer (2013) was adapted to specifically amplify sequences of Ig gene variable regions from single B-cells. Random hexamer primers (Roche) were used to reverse transcribe total B-cell RNA with SuperScript II[®] reverse transcriptase (Invitrogen). Ig gene specific primer mixes (table 4.1) were then used to amplify Ig gene variable regions from cDNA in a nested PCR reaction under the following conditions: 94°C for 15 min followed by 50 cycles of 94°C for 30 s, 58°C (IgH and Ig k) or 60°C (Ig l) for 30 s and 72°C for 55 s followed by 72°C for 10 min. 3.5 μL of first round PCR product was then used in a second round of PCR with the following conditions: 94°C for 15 min followed by 50 cycles at 94°C for 30 s, 62°C for 30 s and 72°C for 45 s followed by 72°C for 10 min. 10 μL of second round PCR product was loaded onto a 2% agarose gel and products of the correct size (450 bp for IgH, 510 bp for Igk and 405 bp for Ig λ) were sequenced using the BigDye Terminator v3.1 Cycle Sequencing kit and analysed as described above (section 4.2.1).

4.2.7 Ig gene sequence analysis

Sequences were aligned using Clustal Omega (Sievers et al., 2011) and then adjusted manually to preserve Chothia numbering (Al-Lazikani et al., 1997). Complementarity-determining regions (CDRs) were determined by alignment to germline sequences in the IMGT database (Giudicelli et al., 2006) using IMGT/V-QUEST (Giudicelli et al., 2004). Alignment to all sequences in the IMGT database was performed with igBLAST (Ye et al., 2013). Unusual residues were identified by comparison to antibody sequences in the European Nucleotide Archive (ENA), Kabat database (Johnson and Wu, 2001) and the Protein Data Bank (PDB) (Berman et al., 2002) by abYsis (Swindells et al., 2017). Estimates of selection on Ig sequences were calculated using Bayesian estimation of Ag-driven SElectioN (BASELINE) with the focused algorithm (Uduman et al., 2011, Yaari et al., 2012).

Reaction	Primer	Sequence
IgH first PCR	5'L-VH1	ACAGGTGCCCCACTCCCAGGTGCAG
	5'L-VH3	AAGGTGTCCAGTGTGARGTGCAG
	5'L-VH4/6	CCCAGATGGGTCTGTCCCAGGTGCAG
	5'L-VH5	CAAGGAGTCTGTTCCGAGGTGCAG
	3'C μ CH outer	GGAAGGAAGTCCTGTGCGAGGC
	3'C γ CH1	GGAAGGTGTGCACGCCGCTGGTC
	3'C α CH1	TGGGAAGTTTCTGGCGGTCACG
IgH second PCR	5'AgeI VH1	CTGCAACCGGTGTACATTCCCAGGTGCAGCTGGTGCAG
	5'AgeIVH1/5	CTGCAACCGGTGTACATTCCGAGGTGCAGCTGGTGCAG
	5'AgeIVH3	CTGCAACCGGTGTACATTCTGAGGTGCAGCTGGTGGAG
	5'AgeIVH3-23	CTGCAACCGGTGTACATTCTGAGGTGCAGCTGTTGGAG
	5'AgeIVH4	CTGCAACCGGTGTACATTCCCAGGTGCAGCTGCAGGAG
	5'AgeIVH4-34	CTGCAACCGGTGTACATTCCCAGGTGCAGCTACAGCAGTG
	3'C μ CH1	GGGAATTCTCACAGGAGACGA
	3'IgG (internal)	GTTCCGGGGAAGTAGTCCTTGAC
	3'C α CH1-2	GTCCGCTTTCGCTCCAGGTCACACT
Igλ first PCR	5'LV λ 1	GGTCCTGGGCCCAGTCTGTGCTG
	5'LV λ 2	GGTCCTGGGCCCAGTCTGCCCTG
	5'LV λ 3	GCTCTGTGACCTCCTATGAGCTG
	5'LV λ 4/5	GGTCTCTCTCSCAGCYGTGCTG
	5'LV λ 6	GTTCTTGGGCCAATTTTATGCTG
	5'LV λ 7	GGTCCAATTCYCAGGCTGTGGTG
	5'LV λ 8	GAGTGGATTCTCAGACTGTGGTG
	3'C λ	CACCAGTGTGGCCTTGTTGGCTTG
Igλ second PCR	5'AgeIV λ 1	CTGCTACCGGTTCTGGGCCCAGTCTGTGCTGACKCAG
	5'AgeIV λ 2	CTGCTACCGGTTCTGGGCCCAGTCTGCCCTGACTCAG
	5'AgeIV λ 3	CTGCTACCGGTTCTGTGACCTCCTATGAGCTGACWCAG
	5'AgeIV λ 4/5	CTGCTACCGGTTCTCTCTCSCAGCYGTGCTGACTCA
	5'AgeIV λ 6	CTGCTACCGGTTCTTGGGCCAATTTTATGCTGACTCAG
	5'AgeIV λ 7/8	CTGCTACCGGTTCCAATTCYCAGRCTGTGGTGACYCAG
	3'XhoIC $\lambda\lambda$	CTCCTCACTCGAGGGYGGGAACAGAGTG
Igκ first PCR	5'LV κ 1/2	ATGAGGSTCCCYGCTCAGCTGCTGG
	5'LV κ 3	CTCTTCCTCTGCTACTCTGGCTCCCAG
	5'LV κ 4	ATTTCTCTGTTGCTCTGGATCTCTG
	3'C κ 543	GTTTCTCGTAGTCTGCTTTGCTCA
Igκ second PCR	5'PanV κ	GTTTCTCGTAGTCTGCTTTGCTCA
	3'C κ s494	GTGCTGTCCTTGCTGTCCTGCT

Table 4.1 List of primer mixes used in amplification of Ig gene variable regions. Primers used for nested PCR amplification of heavy chain (IgH) and light chain (Ig λ and Ig κ) variable regions (Wardemann and Kofer, 2013). All primer sequences shown in 5' to 3'.

4.3 Results

4.3.1 Cysteine residue successfully introduced to T-cell epitope of polyvalent hybrid MSP-1 antigen

PVH antigen F contains no cysteine residues (Tetteh and Conway, 2011) meaning a cysteine residue could be introduced which would then allow reaction of the free sulphydryl with maleimide-biotin resulting in targeted biotinylation. In order to introduce a cysteine residue into PVH antigen F, primers were designed that would anneal to the gene sequence encoding part of the block 1 sequence present in PVH antigen F and introduce a thymine at position 78 resulting in an S26C mutation in the encoded amino acid sequence. The primers were designed to initiate DNA synthesis away from the site of mutation and contained complementary 3' sequences, such that the plasmid encoding PVH antigen F would be linearised and then re-ligated *in vivo* when cloned into *E. coli* cells as described previously (Klock and Lesley, 2009). Following degradation of non-linear plasmid by *DpnI*, an endonuclease that targets methylated DNA, *E. coli* were transfected and selected by resistance to ampicillin. Sequencing of the plasmids present in these colonies showed that four out of 9 had the desired mutation (figure 4.5). Subsequent expression of the PVH antigen F_{S26C} construct was successful (figure 4.6).

		M	V	T	H	E	S	Y	
PVH antigen F	63	atggtgacccatgaaagctat							83
PVH antigen F _{S26C}	63	atggtgacccatgaa t gctat							83
		M	V	T	H	E	C	Y	

Figure 4.5. Sequence alignment of PVH antigen F_{S26C} with PVH antigen F shows introduction of cysteine. Primers designed to introduce an A → T mutation in codon 26 of PVH antigen F were used to linearise pET-15-B plasmids encoding PVH antigen F. Linearised, mutated plasmids were transfected into *E. coli* cells where *in vivo* ligation re-constituted circular plasmids. Inserts of pET-15-B plasmids purified from cloned transfectants were sequenced. The sequence from the plasmid present in colony 6 is shown compared to the parental plasmid, sequenced by the same method, along with the predicted amino acid sequence. The mutated base is high-lighted.

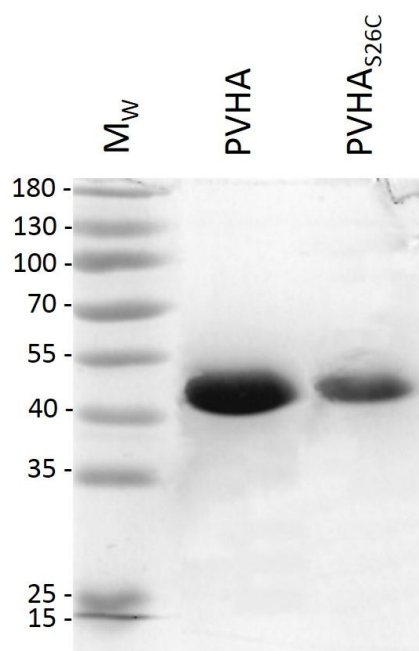


Figure 4.6 Analysis of purified PVH antigens. Following purification of PVH antigens from *E. coli* cell lysate on Ni-NTA beads (ThermoFisher Scientific), 8 μ L samples were loaded onto a 10% SDS-PAGE gel with 5 μ L prestained protein markers (Fermentus). Gel was run at 120 V for 1 hour and then stained (in 0.1% Coomassie R250 (Bio-Rad), 10% acetic acid, 40% methanol) for 1 hour and then de-stained (in 20% methanol, 10% acetic acid) overnight. PVH antigen bands migrated with an apparent mass of 42 kDa (above predicted molecular weight of 27 kDa) as reported previously (Tetteh and Conway, 2011).

4.3.2 Chemical biotinylation and tetramerisation of polyvalent hybrid antigen fails

In order to incorporate a biotin moiety into the PVH antigen F, the free sulphhydryl introduced into the PVH antigen F_{S26C} was reacted with maleimide-PEG₁₁-biotin (comprising a maleimide group joined to biotin by a polyethylene glycol linker). It was expected that this biotinylated PVH antigen (BioPVH antigen F_{S26C}) could then be tetramerised via ligation of single PVH antigens to each of the four biotin binding sites present in streptavidin (SA) tetramers. To assay tetramerisation, BioPVH antigen F_{S26C} was combined with SA in a range of molar ratios and the resulting complexes run on a native protein gel. Formation of multimeric complexes would result in size shift of the protein band, which is not seen (figure 4.7), indicating that tetramers were not formed. The failure of chemical biotinylation of this antigen could result from a local pH that inhibits the formation of the thioether bond between the maleimide and sulphhydryl moieties or steric hindrance preventing access to the sulphhydryl group of the introduced cysteine residue. Biotinylation can also be achieved using the *E.coli* enzyme BirA (Fairhead and Howarth, 2015). This would require the introduction of a BirA biotinylation consensus sequence (Schatz, 1993) into the PVH antigen F and subsequent validation of this antigen. However, for expediency, it was decided that a well validated MSP-1 antigen that had been expressed by a collaborator (section 4.2.2 (Crosnier et al., 2013)) and already included a biotin moiety (introduced by addition of a BirA biotinylation sequence and coexpression with BirA) would be used for production of MSP-1 antigen tetramers.

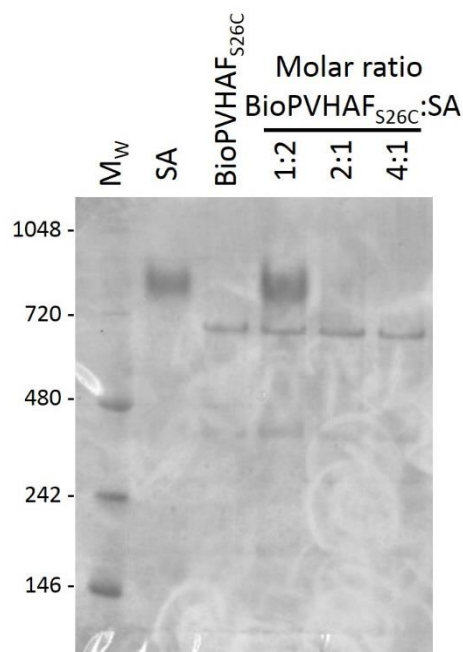


Figure 4.7. Native protein gel showing no change in mass of biotinylated polyvalent hybrid antigen following incubation with streptavidin. A cysteine residue was introduced into the block 1 sequence of PVH antigen F. The free sulphhydryl of the mutant protein (PVH antigen F_{S26C}) was biotinylated using maleimide chemistry. Biotinylated PVH antigen F_{S26C} (BioPVHAF_{S26C}) was mixed with streptavidin (SA, Sigma) at a range of molar ratios and incubated at 4°C for 30 minutes prior to non-denatured samples being run on a 4-12% NuPAGE native acrylamide gel (Novex). NativeMark™ unstained protein markers (M_w , Novex), Streptavidin (SA) and Biotinylated PVH antigen F_{S26C} (BioPVHAF_{S26C}) were loaded in lanes one, two and three respectively. Multimerisation of BioPVHAF_{S26C} through SA binding to biotin would result in a size shift, which is not seen.

4.3.3 Tetramerisation of biotinylated full-length MSP-1 was optimised

A panel of biotinylated merozoite surface proteins had been expressed in mammalian expression systems for use in a high-throughput screen to detect erythrocyte invasion ligands (Bartholdson et al., 2013, Crosnier et al., 2013). This panel included full-length 3D7 MSP-1, which was obtained for use in this study. This antigen is expressed as a fusion with domains three and four of rat CD4. There is no reason to expect that B-cells would recognise this protein, and it has been shown that it is not reactive with a panel of adult human sera (Crosnier et al., 2013). It should be noted that to avoid addition of large glycans during expression that would not be present in native *P. falciparum* proteins, potential N-linked glycosylation sites were mutated in the MSP-1 antigen. This results in 14 single amino acid substitutions (serine or threonine to alanine) in MSP-1. Whilst it is possible that these changes would alter the recognition of these antigens by individual B-cells, it is unlikely that these small changes would have a large impact on overall immunogenicity. Indeed this antigen was shown to be immunoreactive with Kenyan adult sera (Crosnier et al., 2013).

Due to co-expression with BirA and the presence of a C-terminal BirA biotinylation site, MSP1-BLH is expressed with a biotin tag (BioMSP1-BLH, figure 4.8) (Crosnier et al., 2013). This allows for tetramerisation of the MSP-1 protein by binding to SA tetramers. Use of R-pychoerythrin (R-PE) labelled SA (SAPE) would then render this complex fluorescent and allow for the labelling of B-cells recognising MSP-1 (see below section 4.3.4). An excess of BioMSP1-BLH was undesirable as free MSP-1 protein could compete for BCR binding, resulting in lower fluorescence of antigen specific B-cells. Excess of SAPE was also undesirable as it would increase the number of monomeric, dimeric and trimeric BioMSP1-BLH-SAPE complexes; B-cells bind antigen tetramers with a higher affinity than monomers (Franz et al., 2011) and so presence of these smaller complexes could also reduce antigen specific B-cell labelling. Hence, optimisation of the molar ratio of BioMSP1-BLH:SA

tetramers⁴ was performed to identify the ratio at which the majority of BioMSP1-BLH was present in a tetrameric complex with SA (BioMSP1-BLH-SA)₄. Primarily, BioMSP1-BLH was combined with SA in a wide range of molar ratios (1:2 – 64:1) and the resulting protein complexes analysed by native gel electrophoresis showing that the optimal molar ratio was between 8:1 and 16:1 (figure 4.9a). To get a finer estimate of the optimal molar ratio, BioMSP1-BLH was combined with SA at a range of molar ratios between 8:1 and 16:1 (and also at 6:1 and 18:1) and analysed by native gel electrophoresis, revealing that a molar ratio of 14 BioMSP1-BLH to 1 SA tetramer was optimal for formation of BioMSP1-BLH tetramers (figure 4.9b).

⁴ Streptavidin (SA) was used in place of R-pychoerythrin (R-PE) labelled SA (SAPE) as it is a much smaller protein complex and therefore amenable to analysis by native gel electrophoresis. The molar ratio at which biotinylated proteins form tetramers with SA and SAPE is predicted to be very similar as the introduction of the R-PE label does not impact the affinity of SA for biotin (GOTHOT, A., GROSDENT, J. C. & PAULUS, J. M. 1996. A strategy for multiple immunophenotyping by image cytometry: model studies using latex microbeads labeled with seven streptavidin-bound fluorochromes. *Cytometry*, 24, 214-25.).

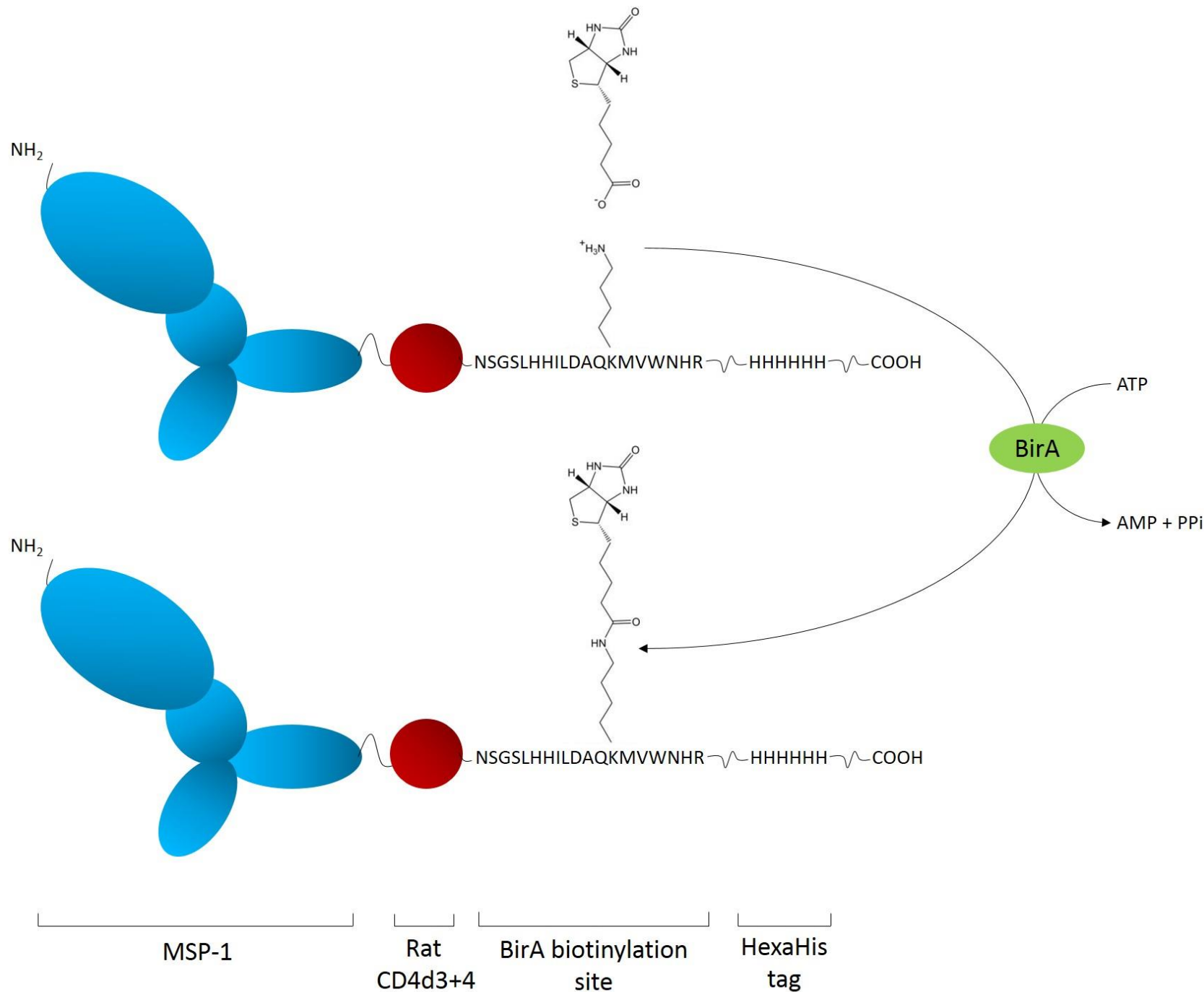


Figure 4.8 Schematic representation of biotinylation of MSP-1 construct. BioMSP1BLH consists of full-length recombinant MSP-1 (3D7 allelic sequence, blue) fused at the C-terminus with domains 3 and 4 of rat CD4 (rat CD4d3+4, red) which has a BirA biotinylation site (sequence shown) followed by a hexa-histidine tag (sequence shown) at the C-terminus. Co-expression with *E. coli* BirA enzyme (green) results in the biotinylation of the lysine residue of the BirA biotinylation site. Figure adapted from (Fairhead and Howarth, 2015)

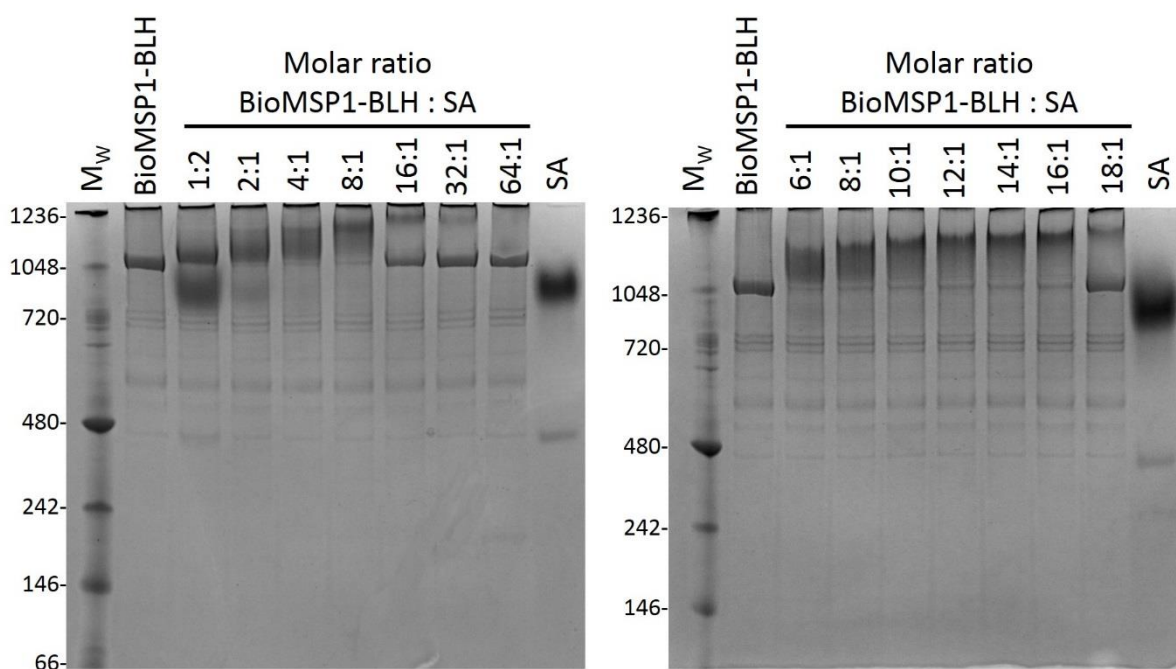


Figure 4.9. Native PAGE showing change in mass of biotinylated MSP-1 following incubation with streptavidin. Biotinylated MSP1-BLH (BioMSP1-BLH) contains a BirA biotinylation site and was co-expressed with BirA resulting in enzymatic biotinylation. BioMSP1-BLH was mixed with a range of molar ratios of streptavidin (SA, lanes 3-9). NativeMark protein markers, BioMSP1-BLH and SA were run for comparison (lanes 1,2 and 10 respectively). Multimerisation of BioMSP1-BLH through binding to SA results in an increase in size shift. At a ratio of 14 BioMSP1-BLH:1 SA (lane 7, right hand gel) the majority of BioMSP1-BLH is present as the largest complex consisting of four molecules of BioMSP1-BLH bound to the SA tetramer. This complex is present at higher ratios of BioMSP1-BLH:SA, but there is greater excess of BioMSP1-BLH.

4.3.4 Isolation of MSP-1 specific memory B-cells

In order to isolate MSP-1 specific memory B-cells for the production of monoclonal antibodies, B-cells were enriched from peripheral blood mononuclear cells (PBMC) collected from adults living in a malaria endemic region. Viability of PBMC, as determined by Trypan blue exclusion, was low for malaria exposed donors (mean 15%, range 3.3-35%) as they had been stored at -80°C for 17-20 months. In comparison the viability was high for the more recently sampled naïve donors (mean 86%, range 71-100%, $p < 0.001$). Antigen specific memory B-cells were sorted based on labelling with anti-CD19 and anti-CD27 antibodies and binding to MSP-1-SAPE tetramers (figure 4.10). CD27⁺ memory B-cells were found to be 32.6% (SD = 11.1%) of circulating B-cells, consistent with frequencies in previously published data (Morbach et al., 2010). One million one hundred and sixty thousand memory B-cells from 16 malaria exposed donors were analysed by flow cytometry (cells from 9 donors were pooled and run as three samples) from which 82 (7 in 100,000) were antigen positive (table 4.2, appendix 7.9). All 82 antigen positive B-cells were isolated by sorting and lysed for amplification and sequencing of mRNA.

In order to determine if labelling with MSP-1-SAPE was due to specific antigen binding, B-cells from 10 naïve donors were prepared under the same conditions as those from exposed donors. These samples had almost identical frequencies of memory B-cells as the exposed samples (31.4%, SD = 8.92%). One million eight hundred and ten thousand memory B-cells were sampled and 70 (4 in 100,000) were identified as antigen positive (table 4.2). Whilst antigen positive memory B-cells occurred at a lower frequency in naïve than in exposed samples, the difference was not statistically significant ($p = 0.19$, Wilcoxon rank sum). However, it is interesting to note that four out of 10 exposed samples had frequencies of antigen positive memory B-cells in excess of 1 in 10,000, whereas only one naïve sample had a frequency greater than this (table 4.2). In both the naïve and exposed samples, there was a large amount of variation in the frequency of antigen positive cells between individuals (figure 4.11).

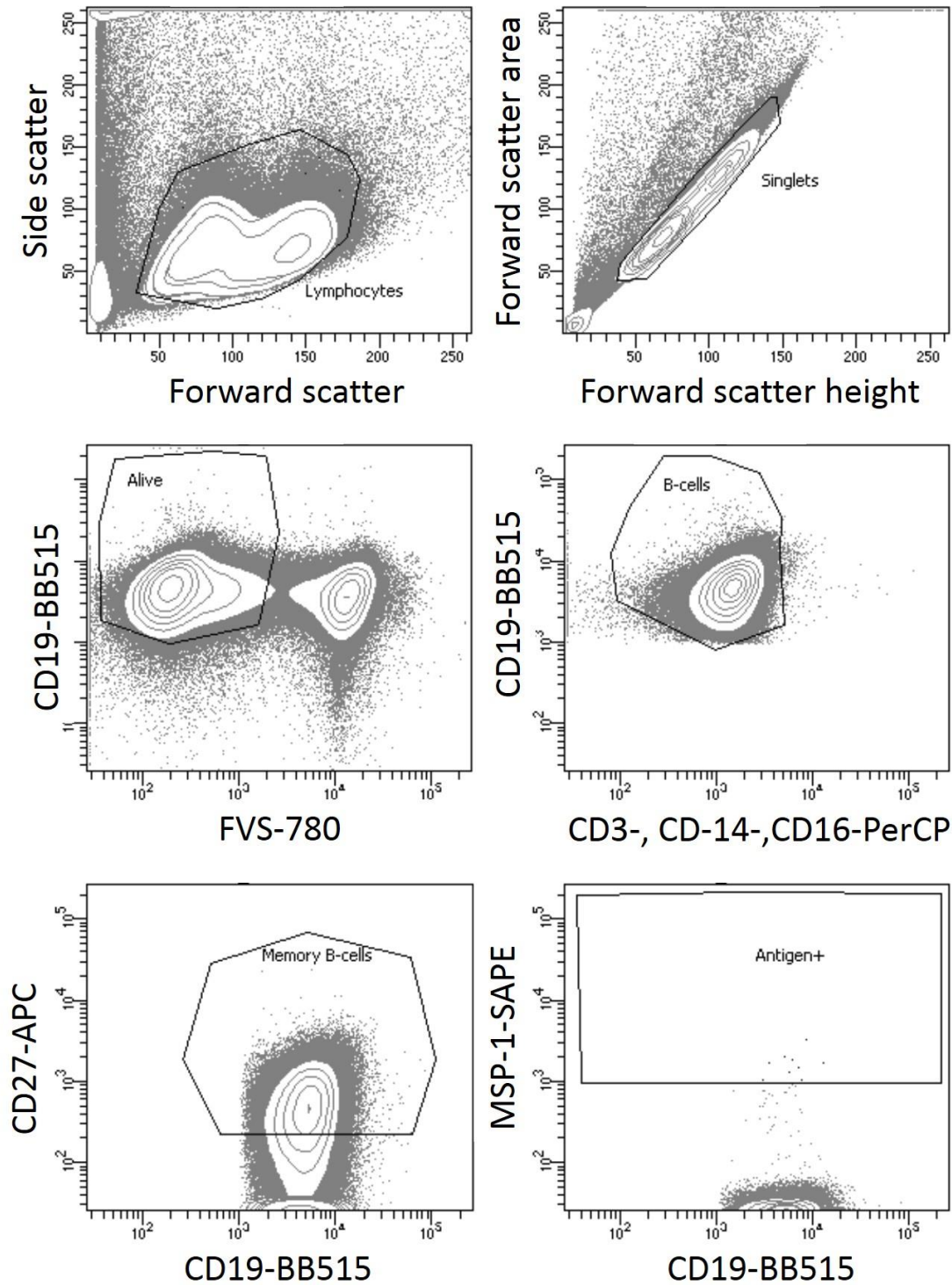


Figure 4.10. Gating strategy for isolation of MSP-1 specific B-cells. Lymphocytes were selected from magnetically enriched B-cells based on forward- and side scatter characteristics. Doublets were excluded by comparison of width and area of forward scatter signal. Dead cells were excluded by increased staining with FVS-780. Non-B-cells were excluded based on expression of CD3, CD14 and CD16. Memory B-cells were selected by expression of CD19 and high expression of CD27. MSP-1-specific cells were isolated on the basis of labelling with MSP-1-SAPE.

	Sample	Memory B-cells (1000s)	Antigen positive	Frequency antigen positive (per 10,000 cells)
naive	503M	43.3	0	0.0
	332M	528	63	1.2
	527F	126	2	0.16
	247F	165	0	0.0
	357F	161	2	0.12
	642F	62	0	0.0
	407F	246	0	0.0
	063F	104	4	0.38
	528M	128	0	0.0
	463M	83.3	4	0.48
	Total	1810	70	0.39
exposed	EIMKB031928	253	62	2.5
	EIMKB32	104	12	1.6
	EIMKB40	40.4	1	0.25
	EIMK024849	484	2	0.041
	EIMKB10	76	0	0.0
	EIMKB05	0.64	1	16
	EIMKB44	87.6	2	0.23
	EIMKB24	68.8	0	0.0
	EIMKB38	3.75	0	0.0
	EIMKB092742	1.81	2	11
	Total	1160	82	0.71

Table 4.2 Cell counts for malaria exposed and naïve samples. B-cells were isolated from adults with no history of malaria (naïve) and adults resident in a malaria endemic region (exposed). Memory B-cells were identified by expression of CD27. Antigen positive cells were identified by binding to MSP-1-SAPE. Cell counts are shown along with the frequency of antigen positive cells per 10,000 memory B-cells. Totals are for each group (naïve and exposed).

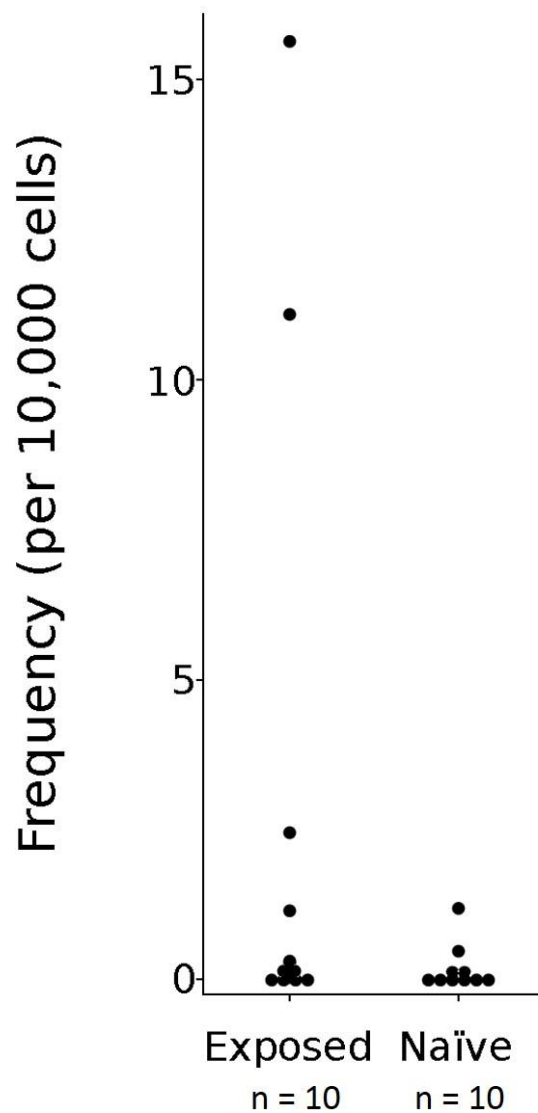


Figure 4.11. Comparison of MSP-1 positive memory B-cells between malaria exposed and naïve individuals. Memory B-cells were incubated with MSP-1-SAPE tetramers. Antigen positive cells were identified by flow cytometry. The frequency of antigen positive cells per 10,000 memory B-cells is shown for malaria exposed and naïve adults. Although a greater number of the malaria-exposed individuals had MSP-1 positive B-cells at frequencies > 1 in 10,000, there was no significant difference between the populations.

4.3.5 Ig gene variable regions sequenced for two antigen positive B-cells

RT-PCR followed by nested PCR was performed on all 82 sorted B-cells from 14 donors in order to amplify the Ig variable regions expressed by the cell as part of the B-cell receptor (BCR). Each sort plate contained a well into which 20 B-cells were sorted to act as a positive control and at least four wells with no cells as a negative control. Three out of five (60%) positive control wells yielded PCR products, however, PCR amplification from non-control wells was seen on both of the plates for which the positive controls failed. There was no amplification from negative control wells (0 out of 22). Six heavy chain and 15 (10 kappa and 5 lambda) light chain genes were successfully amplified from 8 out of 82 (9.7%) antigen positive memory B-cells from 5 donors, as determined by presence of PCR products of the correct size (450 bp for IgH, 510 bp for Igk and 405 bp for Igλ). The frequency of successful amplification of the variable region from B-cell cDNA was much lower than had been observed in preliminary work done in the same laboratory using non-antigen specific memory B-cells from malaria-naïve donors, in which 74 out of 79 (94%) cells yielded at least one PCR product, with 44 (56%) of these having a heavy chain and a single light chain (Serene, 2015).

Two (2.4%) of the 82 antigen positive memory B-cells (EIMKB32 cell B11 and EIMKB031928 cell C11) yielded both a heavy and light chain variable region (figure 4.12). Sequence analysis of these PCR products showed both cells were expressing Ig gamma (IgG1) heavy chains and lambda light chains with productive V(D)J recombinations (figure 4.12). V, D and J gene usage was determined by alignment to V-BASE and Immunogenetics (IGMT) databases of germline sequences. Both BCR sequences had a high number of mutations (86 bases out of 1.16 kb across all four immunoglobulin chains) compared to the reference germline sequences, suggesting a high degree of somatic hypermutation (table 4.3). By comparison to sequences in the Kabat, IMGT and Protein Data Bank (PDB) databases (Bernstein et al., 1977, Johnson and Wu, 2001, Lefranc et al., 2015) it was determined that

the heavy chain of EIMKB32-B11 also encodes a high number (9 out of 98) of unusual residues that occur at that position in less than 1% of known antibody sequences (figure 4.12). Two such residues are also found in each of the two light chains from EIMKB32-B11 and EIMKB031928-C11.

Affinity maturation is a process by which SHM of the BCR variable domains generates a pool of B-cell clones which compete for available antigen with other B-cells and soluble antibody molecules, resulting in selection of BCRs that have increased affinity for their cognate antigen (section 4.1). If there was no selection acting on variants produced by SHM the ratio of synonymous to non-synonymous mutations throughout the variable domain would be determined solely by the mechanism of SHM (Dunn-Walters and Spencer, 1998, Hershberg et al., 2008). As the FR regions form the structure of the variable region, it is predicted that non-synonymous mutations in these regions will tend to be removed during affinity maturation, as they will not increase antigen affinity and may lead to non-functional BCRs (Siskind and Benacerraf, 1969, Hershberg et al., 2008). Indeed the ratio of non-synonymous to synonymous mutations in the 16 FRs of the two BCR sequences analysed here is lower than predicted by a model of SHM without selection ($p < 0.01$), suggesting that non-synonymous changes in these sequences have been selected against in the process of affinity maturation (table 4.4).

A ratio of non-synonymous to synonymous mutations in the CDRs greater than what is predicted under a model of neutral selection is indicative of antigen-driven, positive selection on the immunoglobulin sequence, as the CDRs contain the majority of amino acid residues that are in direct contact with antigen BCRs (Siskind and Benacerraf, 1969, Hershberg et al., 2008). There is an increase in non-synonymous mutations in the CDRs expressed by EIMKB32-B11 BCR above what is predicted by the null model although this is not significant (table 4.4). The ratio of non-synonymous to synonymous mutations in the CDRs expressed by EIMKB031928-C11 is lower than that predicted by the model of neutral selection, although this is not significant (table 4.4).

The sequence of the heavy chain of the BCR of EIMKB32-B11 was found to be highly similar (79.6% amino acid sequence identity) to that of Pf143 (figure 4.13), an antibody binding fragment (Fab), found by screening *E. coli* expressed immunoglobulin genes amplified from six patients being treated for malaria for binding to the C-terminal fragment of MSP-1 (MSP-1₁₉) (Cheng et al., 2007). Surface plasmon resonance demonstrated that this Fab has high affinity for the conserved region of MSP-1₁₉ and immunofluorescence experiments indicated binding to merozoites (Cheng et al., 2007). The sequence of the heavy chain variable region of Pf143 also contains a high number (7) of rare residues found in less than 1% of antibody sequences. Four of these 7 residues, present in the CDR2 and FR3, are shared with the BCR of EIMKB32-B11 (figure 4.13). The CDR3 region of the Ig molecule is usually the most important in determining its affinity (Rock et al., 1994, Xu and Davis, 2000). The heavy chain expressed by EIMKB32-B11 encodes a very long (24 amino acid) CDR3 (figure 4.12), which is divergent from the sequence encoded by Pf143 and all other CDR3 sequences present in IMGT, Kabat and PDB databases.

The PCR strategy used to amplify the variable regions results in the amplification of 99 bp of the IgH constant region. The BCRs of both EIMKB32-B11 and EIMKB031928-C11 have a rare variant in this region that results in the change of amino acid triplet SSK to CSR. This variant is also present in the Pf143 constant region (figure 4.14).

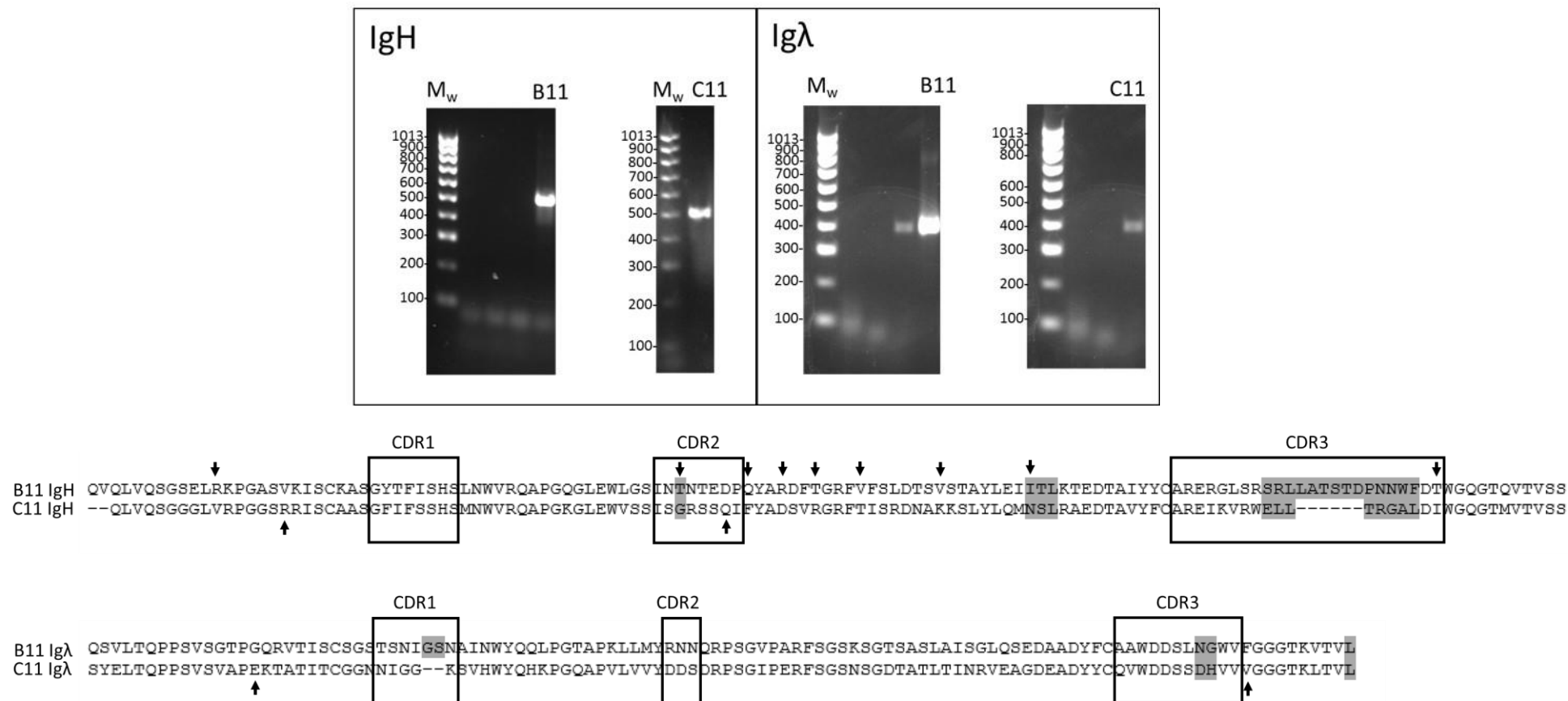


Figure 4.12. Deduced amino acid sequence of Ig variable regions from two MSP-1-specific memory B-cells. Ig gene transcripts were reverse transcribed from two MSP-1-specific memory B-cells and the variable regions were amplified by nested PCR. Agarose gels of second round PCR products, stained with ethidium bromide and visualised with UV light showing amplification of heavy chain (IgH) and lambda light chain (Igλ) from EIMKB32 cell B11 (B11) and EIMKB031928 cell C11 (C11) are shown (top). PCR products were sequenced and the encoded amino-acid sequences of the light and heavy chain variable regions are shown for both BCRs (bottom). Immunogenetics (IMGT® (Giudicelli et al., 2006)) complementarity-determining regions (CDRs), as determined by IgBLAST (Ye et al., 2013), are labelled. Insertions are highlighted in grey. Arrows show positions of unusual residues that occur in fewer than 1% of antibody sequences as determined by abYsis (Swindells et al., 2017).

Antibody	Chain	V gene (reference)	% Identity	D gene (reference)	% Identity	J gene (reference)	% Identity
EIMKB32-B11	IgH	VI-4.1b (Shin et al., 1991)	89.8 (264/294)	D3-3 (Corbett et al., 1997)	100 (7/7)	JH5a (Ravetch et al., 1981)	91.7 (44/48)
	Igλ	1c.10.2 (Williams et al., 1996)	95.6 (281/294)	N/A	N/A	JL3b (Kawasaki et al., 1997)	94.4 (34/36)
EIMKB031928-C11	IgH	WHG16 (Kuppers et al., 1992)	91.7 (264/288)	D1-26 (Corbett et al., 1997)	92.3 (12/13)	JH3b (Mattila et al., 1995)	95.6 (43/45)
	Igλ	V2-14 (Kawasaki et al., 1997)	93.4 (269/288)	N/A	N/A	JL2/JL3a (Udey and Blomberg, 1987)	94.4 (34/36)

Table 4.3 V(D)J gene usage for variable regions of two anti-MSP-1 antibodies. Sequences of variable regions of Ig gene transcripts were aligned to germline sequences in the V-BASE and Immunogenetics (IGMT) databases of germline sequences with IMGT/V-QUEST (Giudicelli et al., 2004) to determine

usage of V, D and J genes. For each variable region sequenced the identity of the germline gene with the closest nucleotide sequence is shown with the percentage identity and number of matching bases.



Figure 4.13 Alignment of EIMKB32-B11 IgH gene sequence with high affinity anti-MSP-119 Fab sequence. Alignment of deduced amino acid sequence of the heavy chain of EIMKB32-B11 BCR with the heavy chain of Pf143, a Fab with high affinity for recombinant MSP-1₁₉ (Cheng et al., 2007), showing similar sequence in complementarity-determining regions (CDRs; shaded red) 1 and 2 and rare amino acid residues (found in <1% of antibody sequences; in red) in CDR2 and framework region (FR; shaded green) 3.

Antibody	Sequence	Observed		Expected		Selection (Σ)					
		CDR		FR		CDR		FR			
		NS	S	NS	S	NS	S	NS	S	CDR	FR
EIMKB32-B11	V _H	5	4	15	6	0.15	0.04	0.58	0.23	-0.0475	-0.378
	V _L	3	1	6	2	0.16	0.02	0.58	0.24	0.461	-0.158
	Combined									0.207	-0.268
EIMKB031928-C11	V _H	5	3	9	7	0.27	0.05	0.49	0.19	-0.756	-0.814**
	V _L	3	0	8	5	0.14	0.03	0.61	0.22	0.0792	-0.469
	Combined									-0.339	-0.642
All sequences combined										-0.066	-0.455**

Table 4.4 Analysis of somatic hyper-mutation for evidence of antigen-driven selection. The Ig variable regions encoded by the two MSP-1 specific memory B-cells for which both a heavy and light chain variable region were successfully amplified were aligned to the closest germline gene sequence to determine mutations likely resulting from somatic hypermutation (SHM). Counts of non-synonymous (NS) and synonymous (S) mutations are shown for complementarity determining regions (CDR) and framework regions (FR) for both the heavy (V_H) and light (V_L) chain variable regions compared to the expected frequency of these mutations predicted by the Bayesian estimation of Ag-driven SElectioN (BASELINE) under neutral selection (Uduman et al., 2011, Yaari et al., 2012). BASELINE then uses the odds ratio between observed and expected NS and S mutations to calculate a selection value (Σ) for CDRs and FRs; values > 0 indicate positive selection and values < 0 indicate negative selection. Selection values for each CDR and FR is shown as well as the combined scores for each immunoglobulin molecule and all four sequences. Significance of selection values, as calculated by Z-test, are indicated (** = p < 0.01).

B11-IgH	ASTKGPSVFPLAPCSRSTS
C11-IgH	ASTKGPSVFPLAPCSRSTS
Pf143-IgH	ASTKGPSVFPLAPCSRSTS
IgH-constant	ASTKGPSVFPLAPSSKSTS

Figure 4.14. Alignment of IgH constant region containing rare variant present in three MSP-1-specific antibodies. The amino acid sequence of the heavy chain constant region of EIMKB32-B11 (B11), EIMKB031928-C11 (C11) and Pf-143 are shown. The two amino acid changes relative to the major allele of the constant region (IgH-constant) are highlighted (black box).

4.4 Discussion

To demonstrate that MSP-1 specific memory B-cells could be isolated from malaria exposed donors, fluorescently labelled antigen was produced. Fluorescent tagging of the antigen via ligation of introduced biotin moieties by R-phycoerythrin conjugated streptavidin not only couples the antigen to the brightest fluorophore available but also results in the formation of antigen tetramers that have been shown to increase affinity for BCRs (Franz et al., 2011). Via introduction of a cysteine residue in the block one sequence present in the PVH antigen F, it was expected that chemical biotinylation could be performed to allow streptavidin-mediated tetramerisation and fluorescent labelling of the PVH antigen F to create a reagent for labelling MSP-1 block 2 reactive memory B-cells, which could be used to produce recombinant monoclonal antibodies against this vaccine candidate. PIPE cloning was successfully used to introduce a single nucleotide change into the PVH antigen F sequence. However, this did not allow for chemical biotinylation and subsequent tetramerisation of the antigen. The most probable explanation for the failure of this approach is inhibition of the reduction of the free sulphydryl by maleimide, either by a local non-neutral pH or steric hindrance. In order to overcome this issue, a *BirA* site could be added to the PVH antigen F; co-expression with *BirA* would then result in high-efficiency of *in vivo* biotinylation. Due to time constraints on this project, this approach was not attempted and as an alternative a full-length, biotinylated MSP-1 antigen was obtained from a collaborator. This full length MSP-1 antigen only contains one K1-like sequence and presents a far greater number of epitopes in addition to block 2. The antigen was used to demonstrate that tetrameric antigens can be used to isolate memory B-cells recognising *P. falciparum* antigens from malaria exposed donors.

The number of memory B-cells analysed for malaria exposed individuals was far lower than expected. This was due to the viability of PBMC after thawing of exposed samples, which was much lower than that reported for this method of cryopreservation (Disis et al., 2006). The fact that the viability of naïve cells, was within the range reported (Disis et al., 2006), suggested that the storage

of these samples at -80°C for up to 17 months longer than the intended 3 month temporary storage resulted in damage to the cells. For future work, cells should be transferred to liquid nitrogen storage, if storage for longer than a few months is needed. For optimum yields, it is recommended that cell sorting is done at the site of collection as this would avoid cryopreservation, but this is not currently possible in malaria endemic areas in West Africa due to the lack of cell sorting facilities.

The large degree of variability in the frequency of MSP-1 specific memory B-cells detected in different individuals can be explained by individual differences in the B-cell repertoire and response to this antigen (Riley, 1996) as well as variation in exposure due to heterogeneity in transmission (Bejon et al., 2011, Bousema et al., 2011, Clark et al., 2008, Drakeley et al., 2005, Kreuels et al., 2008, Mueller et al., 2012, Stewart et al., 2009). The identification of memory B-cells from naïve individuals binding to MSP-1 was not expected. Although the vast majority of these cells were identified in a single individual, cells binding MSP-1 were identified in five out of ten individuals (table 4.2). This could be explained by the presence of cross-reactive BCRs or by non-specific binding of memory B-cells to the MSP-1-SAPE tetramer. Given the variation in the frequency of antigen positive memory B-cells in both naïve and exposed samples, analysis of a greater number of samples will be required to determine if the trend for increased antigen binding in exposed samples observed here is significant.

The number of B-cells for which a heavy and light chain sequence was obtained was an order of magnitude below what has been published for this PCR strategy (Tiller et al., 2008) and from work done in our laboratory with non-antigen-specific memory B-cells (Serene, 2015). This could possibly arise from cells being sorted onto the sides of wells rather than the bottom due to accumulation of salt crystals in the nozzle of the cell sorter. Alternatively, increases in storage and preparation time may have meant that B-cell mRNA was degraded prior to sorting. The fact that only three of five wells into which 20 B-cells were sorted gave PCR products suggest the former explanation, as it

would be expected that increasing the number of cells in the well, and thus total mRNA by a factor of 20 would overcome any issues with mRNA quality.

The two BCR variable region sequences presented here are the first sequences of malaria antigen specific BCRs from West Africa. Both BCRs contain rare variants in the heavy chain constant domain. These rare variants are also found in the constant region of an antibody, possibly of African origin, that has been shown to bind to the C-terminal region of MSP-1 (Cheng et al., 2007). The first of these rare variants introduces an additional cysteine residue into the constant region. It is tempting to speculate that this may alter the pattern of cysteine bond formation in the mature antibody molecule which could allow for enhanced binding of the antibody to its cognate epitope or increased capacity for cross-linking of Fc receptors enhancing downstream immune activation.

The BCR of cell EIMKB32-B11 has an unusual heavy chain sequence that contains 11 rare amino acid substitutions and has a CDR3 that is longer than 90% of those found from sequencing over 37 million BCRs from three adult donors (DeWitt et al., 2016). The CDR3 typically exerts the greatest influence over antigen recognition (Xu and Davis, 2000) and the unusual properties of the CDR3 of EIMKB32-B1 could therefore endear this antibody with interesting properties. Indeed, a recent study discovered an antibody with an insertion resulting from recombination with LAIR-1 that has a broad specificity against variant antigens of the RIFIN family expressed on the surface of *P. falciparum* infected RBCs (Tan et al., 2016). To determine if the unusual CDR3 expressed by EIMKB32-B11 gives this antibody molecule broad specificity against MSP-1 it will be necessary to express this antibody and determine its affinity for a panel of MSP-1 antigens representing the naturally occurring variation seen in this antigen (Miller et al., 1993).

Analysis of the distribution of synonymous and non-synonymous changes in the variable region sequences introduced by somatic hypermutation (SHM) during B-cell maturation indicates that non-synonymous changes in the FRs of both antibody sequences had been subjected to negative selection (table 4.4). This is evidence that both memory B-cells have undergone affinity maturation,

a process in which mutations in the FRs are likely to be selected against as they could result in non-functional BCRs. The same process would result in signatures of positive selection on the CDRs as mutations in these antigen-binding regions that increase antigen affinity would be positively selected for. Although not significant, there is a signature of positive selection on the CDRs of EIMKB32-B11, suggesting that this memory B-cell has undergone affinity maturation (table 4.4). The remarkable sequence similarity with the heavy chain of an antibody fragment, Pf143, that has been shown to have high affinity for MSP-1 (Cheng et al., 2007) might suggest that this memory B-cell does indeed encode an antibody with affinity for MSP-1. However, if the sequence of the Pf143 antibody fragment was amplified from a donor of African descent, this similarity could be the result of shared germline variants that show up as rare mutants due to under representation of African germline Ig gene sequences in Ig gene databases. Unfortunately, amplified Ig gene sequences from eight donors (six of Japanese and two of African descent) were mixed before cloning into expression vectors and selection for affinity to MSP-1₁₉ meaning that the origin of the Pf143 sequence is impossible to ascertain (Professor Hiroshi Tachibana, personal communication). The CDR1 and CDR2 of the IgH chain of EIMKB32-B11 and Pf143 are very similar but the CDR3 and light chains are divergent (figure 4.13). It would therefore be of interest to produce a monoclonal antibody using the BCR sequence from EIMKB32-B11 to investigate how these changes alter antigen specificity.

The CDRs of EIMKB031928-C11 show signatures of negative selection. Whilst not significant, it could indicate that this memory B-cell is the product of polyclonal activation as there is no evidence of positive selection of mutations in the CDRs. PfEMP1, expressed on the surface of infected RBCs, has been reported to stimulate polyclonal expansion of B-cells (Donati et al., 2004), but this has not been reported for MSP-1.

It remains unclear whether the population of memory B-cells in the blood is representative of the cells that will differentiate into antibody producing plasma cells on re-exposure to antigen, and thus contribute to control of future infections (Gourley et al., 2004, Tarlinton, 2006). Evidence from

studies in mice would suggest that migration of memory B-cells from the germinal centres is driven by selection, as the degree of somatic hypermutation seen in circulating B-cells increases with time following antigen exposure (Blink et al., 2005). This could mean that memory B-cells with a higher affinity for antigen are retained in the germinal centres and may not be sampled by isolating memory B-cells from blood but would contribute to future antibody responses. In order to avoid this potential bias, sampling the blood from concurrently infected, clinically immune individuals would allow for the analysis of the plasma cells that are producing antibody in response to infection. Whilst this would require the use of fresh blood, as these cells cannot be cryopreserved, the expansion of this cell population in response to infection should enhance isolation of antigen-specific cells (Franz et al., 2011). In mice infected with *P. chabaudi*, the numbers of MSP-1 specific memory B-cells and antibody secreting plasma cells varied dramatically over the time following infection, suggesting that the timing of sampling will dramatically impact the success of this approach (Nduati et al., 2010).

In conclusion, the work presented here highlights several key areas which need to be addressed in order for labelling with *P. falciparum* antigen tetramers for isolation specific antibody producing cells for use in the production of large numbers of monoclonal antibodies. In order for this approach to yield the best results, fresh blood from infected individuals should be used to allow for the isolation of plasma cells. This will not only boost the yield of monoclonal antibodies but also capture the antibody specificities that are produced in response to infection.

Chapter 5 - Discussion

5.1 The use of short read data for the analysis of population-wide variation in repeat sequences

This thesis has detailed the development of bioinformatic tools that allow short read data to be used for the analysis of polymorphic and repetitive sequences in the context of *P. falciparum* vaccine design. Polymorphic repeat sequences, a common feature of *P. falciparum* antigens (Anders et al., 1988, Feng et al., 2006), can often be grouped into allelic types (Anders et al., 1993). Whilst there is variation in repeat sequence and length within these groups, it is of use to be able to classify individual alleles by this method. This has classically been done by using type-specific PCR to amplify parasite DNA (Farnert et al., 2001). Work presented in this thesis demonstrates that the short read sequence data generated from whole genome sequencing can be leveraged to determine the allelic types present in an isolate. Using MSP-1 block 2 as an example, a sequence library was constructed and validated. Short read sequences from the Pf3K project were then aligned to this sequence library to determine the allelic types present in the isolate. This novel method produced results that agree with historical data for MSP-1 block 2 determined by PCR genotyping as expected given the temporal stability of allele frequencies (Tanabe et al., 2007a, Noranate et al., 2009, Silva et al., 2000). This approach allows for population-wide survey of the alleles present at polymorphic loci of interest, which is of use both in the design of multi-allelic vaccines and in the monitoring of the effects such vaccines would have on allele frequencies in parasite populations.

De novo assembly is known to struggle with repeat sequences (Li et al., 2010). Work presented in this thesis demonstrates that an optimised *de novo* assembly algorithm (Zerbino and Birney, 2008) can assemble *msh1* block 2 repeat sequences. The probability of assembly is dependent on the length of repeat sequence and on the coverage depth (figures 2.4 and 2.5), which can be boosted by first mapping reads to a sequence library. The assembly of reads aligned to a sequence library resulted in generation of the largest single dataset of *msh1* block 2 sequences. However, this

approach will run into difficulties when repeat sequences are longer than the read length, which is the case for a considerable number of *P. falciparum* genes (Aspeling-Jones, 2013). Two algorithms for assembling retroviral genomes use a stepwise approach to extend contigs in which the distance between paired reads is used to confirm the mapping of a reads overlapping the end of a contig (Hunt et al., 2015, Ruby et al., 2013). It would be of interest to see if this approach could be adapted to overcome the issue of long repeat sequences.

The recent advance of long (upwards of 10 kb) read whole genome sequencing technologies, notably Pacific Biosciences (PacBio) and Oxford Nanopore Technologies (ONT) presents another method for overcoming assembly of repeat sequences (Koren and Phillippy, 2015). ONT applications still struggle with the AT rich genome of *P. falciparum* but PacBio sequencing has been used to successfully assemble a complete genome for the 3D7 lab strain (Vembar et al., 2016). However, the increased cost of this sequencing technology combined with the requirement for large amounts of DNA may limit its application to studies at the population level. Whilst the Pf3k consortium has used PacBio sequencing to look at a small number of clinical isolates, the real power of this technology may be in the rapid assembly of novel genomes, as has been done for *P. malariae* and *P. ovale* (Rutledge et al., 2016).

The work presented here on *msh1* block 2 is proof of principle that, through alignment to a sequence library, short read data can be used for both typing and assembling highly polymorphic regions. Such approaches could be applied to any gene of interest, so long as sufficient sequence data is available to construct a sequence library. For well-studied antigens, sufficient sequence data is available to construct such libraries, however, this is not the case for novel vaccine candidates that are less well characterised. It is probable that the sequences obtained from *de novo* assembly of read pairs in which at least one mate maps to the polymorphic repeat locus in the 3D7 reference sequence (as done for *msh1* block 2, see section 2.3.2) would be sufficient for the construction of a sequence library, however, this remains to be tested. Given the relatively small number of

sequences required, reference libraries allowing the analysis of population-wide variation of novel repeat antigens could be constructed from long read sequencing of a handful of isolates and lab lines.

Whilst a combination of alignment and *de novo* assembly can yield full length sequence data from short reads, reads from a large fraction (24%) of isolates could not be assembled by this method and this approach does not guarantee that all sequences present in an isolate will be assembled. Due to the fact that intrinsically disordered protein domains, such as those often encoded by repetitive sequences, present linear B-cell epitopes the complete sequence is not necessary for the identification of potential epitopes. This thesis demonstrates that it is possible to extract a the major potential epitope sequences from reads aligned to a sequence library and describes the development of an algorithm that can use these short amino acid sequences and their frequency in the population to design hybrid antigens. Many established and novel *P. falciparum* candidate vaccine antigens contain intrinsically disordered domains that are predicted to present linear B-cell epitopes (Feng et al., 2006, Guy et al., 2015) and this approach can be used to design polyvalent hybrid antigens for each of these using existing short read sequence data.

5.2 The use of tetramerised antigens for the isolation of antigen-specific memory B-cells

The ability to isolate immune cells recognising *P. falciparum* antigens promises to allow for both the *ex vivo* analysis of the immune response to such antigens and the production of human monoclonal antibodies for analysis of the mechanisms by which these antibodies control parasite replication. One approach to isolating antigen-specific B-cells employs fluorescently labelled antigen tetramers for use in sorting B-cells. These tetramers have the potential to bind to cognate B-cells with high specificity via the B-cell receptor (BCR) (Franz et al., 2011). This thesis details the production of *P. falciparum* antigen tetramers based on MSP-1 and their use to isolate memory B-cells from malaria exposed adults.

The work presented here suggests that utilisation of the *E. coli* BirA system for biotinylation of antigens is preferable. Previous work has shown that memory B-cells can be isolated from naturally exposed individuals (Muellenbeck et al., 2013) but the work presented in this thesis is the first attempt to do this with tetramerised antigens, which should increase the specificity of B-cell labelling. Whilst this approach was able to isolate memory B-cells, the yield was far lower than expected. This was due to prolonged storage of B-cells at -80°C which limited the number of cells that could be analysed. The successful sequencing of Ig variable regions from two isolated, antigen specific, memory B-cells demonstrates the robustness of the method described by Wardemann and Koffer (2013) and hints at the potential presence of distinct germline Ig genes amongst African populations that have not been well sampled previously. Further work is still needed to prove that the antibody sequences isolated by the approach described here do in fact bind to both recombinant and native MSP-1 protein.

Production of recombinant human monoclonals from B-cells isolated on the basis of binding to fluorescently labelled antigen merits further application to the study of naturally acquired immune responses to *P. falciparum*. This approach avoids any bottleneck introduced by either Epstein Barr Virus (EBV) transformation of B-cells or successful expression of FAb fragments in *E. coli*, which may select against certain subtypes of cells or antibodies. The approach also allows the study of antibody secreting cells *ex vivo*, with the potential to isolate antibody sequences from the plasma cells that actually produce antibodies in response to infection (Franz et al., 2011).

5.3 Tools for the design and validation of vaccine antigens based on polymorphic repeat sequences

Using the methods described in this thesis, polyvalent hybrid antigens can be easily designed for a large panel of antigen domains predicted to present linear B-cell epitopes. These synthetic antigens can be expressed with BirA tags allowing enzymatic biotinylation and subsequent tetramerisation with fluorescently labelled streptavidin. These tetramers can be used to isolate memory B-cells

recognising the hybrid antigens from clinically immune individuals. Expression of monoclonal antibodies from these individuals will then allow analysis of which proposed hybrid antigens are targets of functional immunological memory through assaying the antigen specific monoclonal antibodies for activity in growth inhibition, opsonic phagocytosis and complement-mediated inhibition assays. Combining this functional screening with epitope mapping will enable determination of any bias both in the recognition of linear epitopes and in the functionality of immune responses to them. This information can then be used to inform the design of multi-allelic, multi-antigen vaccines with the potential to elicit memory responses against functional epitopes present in a number of different alleles of different blood stage antigens.

5.4 Concluding remarks

This thesis details the use of short read data for the analysis of highly polymorphic repeat sequence. It is shown that this data can be used for the analysis of repeat sequences, which is of high relevance to the study of *P. falciparum* antigens. The *msp1* block 2 repeat sequence was used in this study as it is very well characterised, nevertheless, the approaches employed here have produced the largest global survey of *msp1* block 2 sequences. Whether antibody responses to MSP-1 block 2 antigens are protective is still an issue of debate. The bioinformatics approaches presented in this thesis could easily be adapted for use with other polymorphic repeat sequences, which are a common feature of *P. falciparum* antigens.

This thesis also outlines the first use of antigen tetramers for the isolation of memory B-cells specific for *P. falciparum* antigens from naturally exposed individuals. Although the approach did not yield the as many cells as expected, it shows that this technique can be used for isolation of memory B-cells from naturally exposed individuals. The work also produced two antibody sequences that are likely to encode antibodies against MSP-1. Future work combining the two methods developed here could produce a library of polyvalent hybrid antigens and cognate human monoclonals that would guide the development of vaccines targeting a number of *P. falciparum* antigens.

6. References

- ACOSTA, C. J., GALINDO, C. M., SCHELLENBERG, D., APONTE, J. J., KAHIGWA, E., URASSA, H., SCHELLENBERG, J. R., MASANJA, H., HAYES, R., KITUA, A. Y., LWILLA, F., MSHINDA, H., MENENDEZ, C., TANNER, M. & ALONSO, P. L. 1999. Evaluation of the SPf66 vaccine for malaria control when delivered through the EPI scheme in Tanzania. *Trop Med Int Health*, 4, 368-76.
- ADAMS, J. H., BLAIR, P. L., KANEKO, O. & PETERSON, D. S. 2001. An expanding ebl family of *Plasmodium falciparum*. *Trends Parasitol*, 17, 297-9.
- ADAMS, J. H., SIM, B. K., DOLAN, S. A., FANG, X., KASLOW, D. C. & MILLER, L. H. 1992. A family of erythrocyte binding proteins of malaria parasites. *Proc Natl Acad Sci U S A*, 89, 7085-9.
- AL-LAZIKANI, B., LESK, A. M. & CHOTHIA, C. 1997. Standard conformations for the canonical structures of immunoglobulins. *J Mol Biol*, 273, 927-48.
- AL-YAMAN, F., GENTON, B., KRAMER, K. J., CHANG, S. P., HUI, G. S., BAISOR, M. & ALPERS, M. P. 1996. Assessment of the role of naturally acquired antibody levels to *Plasmodium falciparum* merozoite surface protein-1 in protecting Papua New Guinean children from malaria morbidity. *Am J Trop Med Hyg*, 54, 443-8.
- ALONSO, P. L., SMITH, T., SCHELLENBERG, J. R., MASANJA, H., MWANKUSYE, S., URASSA, H., BASTOS DE AZEVEDO, I., CHONGELA, J., KOBERO, S., MENENDEZ, C. & ET AL. 1994. Randomised trial of efficacy of SPf66 vaccine against *Plasmodium falciparum* malaria in children in southern Tanzania. *Lancet*, 344, 1175-81.
- ALVES, F. P., GIL, L. H., MARRELLI, M. T., RIBOLLA, P. E., CAMARGO, E. P. & DA SILVA, L. H. 2005. Asymptomatic carriers of *Plasmodium* spp. as infection source for malaria vector mosquitoes in the Brazilian Amazon. *J Med Entomol*, 42, 777-9.
- AMAMBUA-NGWA, A., TETTEH, K. K., MANSKE, M., GOMEZ-ESCOBAR, N., STEWART, L. B., DEERHAKE, M. E., CHEESEMAN, I. H., NEWBOLD, C. I., HOLDER, A. A., KNUEPFER, E., JANHA, O., JALLOW, M., CAMPINO, S., MACINNIS, B., KWIATKOWSKI, D. P. & CONWAY, D. J. 2012a. Population genomic scan for candidate signatures of balancing selection to guide antigen characterization in malaria parasites. *PLoS Genet*, 8, e1002992.
- AMAMBUA-NGWA, A., TETTEH, K. K. A., MANSKE, M., GOMEZ-ESCOBAR, N., STEWART, L. B., DEERHAKE, M. E., CHEESEMAN, I. H., NEWBOLD, C. I., HOLDER, A. A., KNUEPFER, E., JANHA, O., JALLOW, M., CAMPINO, S., MACINNIS, B., KWIATKOWSKI, D. P. & CONWAY, D. J. 2012b. Population Genomic Scan for Candidate Signatures of Balancing Selection to Guide Antigen Characterization in Malaria Parasites. *PLoS Genet*, 8, e1002992.
- AMBROGGIO, X., JIANG, L., AEBIG, J., OBIKOR, H., LUKSZO, J. & NARUM, D. L. 2013. The epitope of monoclonal antibodies blocking erythrocyte invasion by *Plasmodium falciparum* map to the dimerization and receptor glycan binding sites of EBA-175. *PLoS One*, 8, e56326.
- AMPOFO, W. K., BAYLOR, N., COBEY, S., COX, N. J., DAVES, S., EDWARDS, S., FERGUSON, N., GROHMANN, G., HAY, A., KATZ, J., KULLABUTR, K., LAMBERT, L., LEVANDOWSKI, R., MISHRA, A. C., MONTO, A., SIQUEIRA, M., TASHIRO, M., WADDELL, A. L., WAIRAGKAR, N., WOOD, J., ZAMBON, M. & ZHANG, W. 2012. Improving influenza vaccine virus selection: report of a WHO informal consultation held at WHO headquarters, Geneva, Switzerland, 14-16 June 2010. *Influenza Other Respir Viruses*, 6, 142-52, e1-5.
- AMPOMAH, P., STEVENSON, L., OFORI, M. F., BARFOD, L. & HVIID, L. 2014. Kinetics of B cell responses to *Plasmodium falciparum* erythrocyte membrane protein 1 in Ghanaian women naturally exposed to malaria parasites. *J Immunol*, 192, 5236-44.
- ANDERS, R. F., COPPEL, R. L., BROWN, G. V. & KEMP, D. J. 1988. Antigens with repeated amino acid sequences from the asexual blood stages of *Plasmodium falciparum*. *Progress in Allergy*, 41, 148-172.
- ANDERS, R. F., MCCOLL, D. J. & COPPEL, R. L. 1993. Molecular variation in *Plasmodium falciparum*: Polymorphic antigens of asexual erythrocytic stages. *Acta Tropica*, 53, 239-253.

- ANDERSON, T. J., HAUBOLD, B., WILLIAMS, J. T., ESTRADA-FRANCO, J. G., RICHARDSON, L., MOLLINEDO, R., BOCKARIE, M., MOKILI, J., MHARAKURWA, S., FRENCH, N., WHITWORTH, J., VELEZ, I. D., BROCKMAN, A. H., NOSTEN, F., FERREIRA, M. U. & DAY, K. P. 2000. Microsatellite markers reveal a spectrum of population structures in the malaria parasite *Plasmodium falciparum*. *Mol Biol Evol*, 17, 1467-82.
- ANDERSSON, A. C., RESENDE, M., SALANTI, A., NIELSEN, M. A. & HOLST, P. J. 2017. Novel adenovirus encoded virus-like particles displaying the placental malaria associated VAR2CSA antigen. *Vaccine*, 35, 1140-1147.
- ANDRE, F. E. 2003. Vaccinology: past achievements, present roadblocks and future promises. *Vaccine*, 21, 593-5.
- APINJOH, T. O., TATA, R. B., ANCHANG-KIMBI, J. K., CHI, H. F., FON, E. M., MUGRI, R. N., TANGO, D. A., NYINGCHU, R. V., GHOGOMU, S. M., NKUO-AKENJI, T. & ACHIDI, E. A. 2015. *Plasmodium falciparum* merozoite surface protein 1 block 2 gene polymorphism in field isolates along the slope of mount Cameroon: a cross - sectional study. *BMC Infect Dis*, 15, 309.
- ARTAVANIS-TSAKONAS, K., TONGREN, J. E. & RILEY, E. M. 2003. The war between the malaria parasite and the immune system: immunity, immunoregulation and immunopathology. *Clinical & Experimental Immunology*, 133, 145-152.
- ASPELING-JONES, H. 2013. *Repeats in Plasmodium falciparum coding sequences in relation to function and selection*. Molecular Biology of Infectious Disease, London School of Hygiene and Tropical Medicine.
- ASSEFA, S. A., PRESTON, M. D., CAMPINO, S., OCHOLLA, H., SUTHERLAND, C. J. & CLARK, T. G. 2014. estMOL: estimating multiplicity of infection using parasite deep sequencing data. *Bioinformatics*, 30, 1292-1294.
- ASTAGNEAU, P., CHOUGNET, C., LEPERS, J. P., DANIELLE, M., ANDRIAMANGATIANA-RASON, M. D. & DELORON, P. 1994a. Antibodies to the 4-mer repeat of the ring-infected erythrocyte surface antigen (Pf155/RESA) protect against *Plasmodium falciparum* malaria. *Int J Epidemiol*, 23, 169-75.
- ASTAGNEAU, P., ROBERTS, J. M., STEKETEE, R. W., WIRIMA, J. J., LEPERS, J. P. & DELORON, P. 1995. Antibodies to a *Plasmodium falciparum* blood-stage antigen as a tool for predicting the protection levels of two malaria-exposed populations. *Am J Trop Med Hyg*, 53, 23-8.
- ASTAGNEAU, P., STEKETEE, R. W., WIRIMA, J. J., KHOROMANA, C. O. & MILLET, P. 1994b. Antibodies to ring-infected erythrocyte surface antigen (Pf155/RESA) protect against *P. falciparum* parasitemia in highly exposed multigravida women in Malawi. *Acta Trop*, 57, 317-25.
- ATAIDE, R., HASANG, W., WILSON, D. W., BEESON, J. G., MWAPASA, V., MOLYNEUX, M. E., MESHNICK, S. R. & ROGERSON, S. J. 2010. Using an improved phagocytosis assay to evaluate the effect of HIV on specific antibodies to pregnancy-associated malaria. *PLoS One*, 5, e10807.
- AUCAN, C., TRAORE, Y., TALL, F., NACRO, B., TRAORE-LEROUX, T., FUMOUX, F. & RIHET, P. 2000. High immunoglobulin G2 (IgG2) and low IgG4 levels are associated with human resistance to *Plasmodium falciparum* malaria. *Infect Immun*, 68, 1252-8.
- AUDRAN, R., CACHAT, M., LURATI, F., SOE, S., LEROY, O., CORRADIN, G., DRUILHE, P. & SPERTINI, F. 2005. Phase I malaria vaccine trial with a long synthetic peptide derived from the merozoite surface protein 3 antigen. *Infect Immun*, 73, 8017-26.
- AURRECOECHEA, C., BRESTELLI, J., BRUNK, B. P., DOMMER, J., FISCHER, S., GAJRIA, B., GAO, X., GINGLE, A., GRANT, G., HARB, O. S., HEIGES, M., INNAMORATO, F., IODICE, J., KISSINGER, J. C., KRAEMER, E., LI, W., MILLER, J. A., NAYAK, V., PENNINGTON, C., PINNEY, D. F., ROOS, D. S., ROSS, C., STOECKERT, C. J., JR., TREATMAN, C. & WANG, H. 2009. PlasmoDB: a functional genomic database for malaria parasites. *Nucleic Acids Res*, 37, D539-43.
- AWANDARE, G. A., SPADAFORA, C., MOCH, J. K., DUTTA, S., HAYNES, J. D. & STOUT, J. A. 2011. *Plasmodium falciparum* field isolates use complement receptor 1 (CR1) as a receptor for invasion of erythrocytes. *Mol Biochem Parasitol*, 177, 57-60.

- AYI, K., TURRINI, F., PIGA, A. & ARESE, P. 2004. Enhanced phagocytosis of ring-parasitized mutant erythrocytes: a common mechanism that may explain protection against falciparum malaria in sickle trait and beta-thalassemia trait. *Blood*, 104, 3364-71.
- BADIANE, A. S., BEI, A. K., AHOUIDI, A. D., PATEL, S. D., SALINAS, N., NDIAYE, D., SARR, O., NDIR, O., TOLIA, N. H., MBOUP, S. & DURAISINGH, M. T. 2013. Inhibitory humoral responses to the Plasmodium falciparum vaccine candidate EBA-175 are independent of the erythrocyte invasion pathway. *Clin Vaccine Immunol*, 20, 1238-45.
- BAI, T., BECKER, M., GUPTA, A., STRIKE, P., MURPHY, V. J., ANDERS, R. F. & BATCHELOR, A. H. 2005. Structure of AMA1 from Plasmodium falciparum reveals a clustering of polymorphisms that surround a conserved hydrophobic pocket. *Proc Natl Acad Sci U S A*, 102, 12736-41.
- BALDWIN, M. R., LI, X., HANADA, T., LIU, S. C. & CHISHTI, A. H. 2015. Merozoite surface protein 1 recognition of host glycophorin A mediates malaria parasite invasion of red blood cells. *Blood*, 125, 2704-11.
- BALLOU, W. R., BLOOD, J., CHONGSUPHAJAISSIDHI, T., GORDON, D. M., HEPPNER, D. G., KYLE, D. E., LUXEMBURGER, C., NOSTEN, F., SADOFF, J. C., SINGHASIVANON, P. & ET AL. 1995. Field trials of an asexual blood stage malaria vaccine: studies of the synthetic peptide polymer SPf66 in Thailand and the analytic plan for a phase IIb efficacy study. *Parasitology*, 110 Suppl, S25-36.
- BANIC, D. M., DE OLIVEIRA-FERREIRA, J., PRATT-RICCIO, L. R., CONSEIL, V., GONCALVES, D., FIALHO, R. R., GRAS-MASSÉ, H., DANIEL-RIBEIRO, C. T. & CAMUS, D. 1998. Immune response and lack of immune response to Plasmodium falciparum P126 antigen and its amino-terminal repeat in malaria-infected humans. *Am J Trop Med Hyg*, 58, 768-74.
- BANNISTER, L. & MITCHELL, G. 2003. The ins, outs and roundabouts of malaria. *Trends Parasitol*, 19, 209-13.
- BARFOD, L., BERNASCONI, N. L., DAHLBACK, M., JARROSSAY, D., ANDERSEN, P. H., SALANTI, A., OFORI, M. F., TURNER, L., RESENDE, M., NIELSEN, M. A., THEANDER, T. G., SALLUSTO, F., LANZAVECCHIA, A. & HVIID, L. 2007. Human pregnancy-associated malaria-specific B cells target polymorphic, conformational epitopes in VAR2CSA. *Mol Microbiol*, 63, 335-47.
- BARRY, A. E. & ARNOTT, A. 2014. Strategies for designing and monitoring malaria vaccines targeting diverse antigens. *Front Immunol*, 5, 359.
- BARRY, A. E., SCHULTZ, L., BUCKEE, C. O. & REEDER, J. C. 2009. Contrasting population structures of the genes encoding ten leading vaccine-candidate antigens of the human malaria parasite, Plasmodium falciparum. *PLoS One*, 4, e8497.
- BARTHOLDSON, S. J., CROSNIER, C., BUSTAMANTE, L. Y., RAYNER, J. C. & WRIGHT, G. J. 2013. Identifying novel Plasmodium falciparum erythrocyte invasion receptors using systematic extracellular protein interaction screens. *Cellular Microbiology*, 15, 1304-1312.
- BAUM, J., CHEN, L., HEALER, J., LOPATICKI, S., BOYLE, M., TRIGLIA, T., EHLGEN, F., RALPH, S. A., BEESON, J. G. & COWMAN, A. F. 2009. Reticulocyte-binding protein homologue 5 - an essential adhesin involved in invasion of human erythrocytes by Plasmodium falciparum. *Int J Parasitol*, 39, 371-80.
- BEESON, J. G., DREW, D. R., BOYLE, M. J., FENG, G., FOWKES, F. J. & RICHARDS, J. S. 2016. Merozoite surface proteins in red blood cell invasion, immunity and vaccines against malaria. *FEMS Microbiol Rev*, 40, 343-72.
- BEI, A. K. & DURAISINGH, M. T. 2012. Functional analysis of erythrocyte determinants of Plasmodium infection. *Int J Parasitol*, 42, 575-82.
- BEJON, P., COOK, J., BERGMANN-LEITNER, E., OLOTU, A., LUSINGU, J., MWACHARO, J., VEKEMANS, J., NJUGUNA, P., LEACH, A., LIEVENS, M., DUTTA, S., VON SEIDLEIN, L., SAVARESE, B., VILLAFANA, T., LEMNGE, M. M., COHEN, J., MARSH, K., CORRAN, P. H., ANGOV, E., RILEY, E. M. & DRAKELEY, C. J. 2011. Effect of the pre-erythrocytic candidate malaria vaccine RTS,S/AS01E on blood stage immunity in young children. *J Infect Dis*, 204, 9-18.
- BELARD, S., ISSIFO, S., HOUNKPATIN, A. B., SCHAUMBURG, F., NGOA, U. A., ESEN, M., FENDEL, R., DE SALAZAR, P. M., MURBETH, R. E., MILLIGAN, P., IMBAULT, N., IMOUKHUEDE, E. B.,

- THEISEN, M., JEPSEN, S., NOOR, R. A., OKECH, B., KREMSNER, P. G. & MORDMULLER, B. 2011. A randomized controlled phase Ib trial of the malaria vaccine candidate GMZ2 in African children. *PLoS One*, 6, e22525.
- BENGTTSSON, A., JOERGENSEN, L., RASK, T. S., OLSEN, R. W., ANDERSEN, M. A., TURNER, L., THEANDER, T. G., HVIID, L., HIGGINS, M. K., CRAIG, A., BROWN, A. & JENSEN, A. T. 2013. A novel domain cassette identifies Plasmodium falciparum PfEMP1 proteins binding ICAM-1 and is a target of cross-reactive, adhesion-inhibitory antibodies. *J Immunol*, 190, 240-9.
- BENSON, D. A., CAVANAUGH, M., CLARK, K., KARSCH-MIZRACHI, I., LIPMAN, D. J., OSTELL, J. & SAYERS, E. W. 2013. GenBank. *Nucleic Acids Res*, 41, D36-42.
- BERGMANN-LEITNER, E. S., DUNCAN, E. H., MULLEN, G. E., BURGE, J. R., KHAN, F., LONG, C. A., ANGOV, E. & LYON, J. A. 2006. Critical evaluation of different methods for measuring the functional activity of antibodies against malaria blood stage antigens. *Am J Trop Med Hyg*, 75, 437-42.
- BERMAN, H. M., BATTISTUZZI, T., BHAT, T. N., BLUHM, W. F., BOURNE, P. E., BURKHARDT, K., FENG, Z., GILLILAND, G. L., IYPE, L., JAIN, S., FAGAN, P., MARVIN, J., PADILLA, D., RAVICHANDRAN, V., SCHNEIDER, B., THANKI, N., WEISSIG, H., WESTBROOK, J. D. & ZARDECKI, C. 2002. The Protein Data Bank. *Acta Crystallogr D Biol Crystallogr*, 58, 899-907.
- BERNSTEIN, F. C., KOETZLE, T. F., WILLIAMS, G. J., MEYER, E. F., JR., BRICE, M. D., RODGERS, J. R., KENNARD, O., SHIMANOUCHI, T. & TASUMI, M. 1977. The Protein Data Bank. A computer-based archival file for macromolecular structures. *Eur J Biochem*, 80, 319-24.
- BERTANI, G. 1951. Studies on lysogeny. I. The mode of phage liberation by lysogenic Escherichia coli. *J Bacteriol*, 62, 293-300.
- BERZINS, K., PERLMANN, H., UDOMSANGPETCH, R., WAHLIN, B., WAHLGREN, M., TROYE-BLOMBERG, M., CARLSSON, J., BJORKMAN, A. & PERLMANN, P. 1985. Pf 155, a candidate for a blood stage vaccine in Plasmodium falciparum malaria. *Dev Biol Stand*, 62, 99-106.
- BERZINS, K., PERLMANN, H., WAHLIN, B., CARLSSON, J., WAHLGREN, M., UDOMSANGPETCH, R., BJORKMAN, A., PATARROYO, M. E. & PERLMANN, P. 1986. Rabbit and human antibodies to a repeated amino acid sequence of a Plasmodium falciparum antigen, Pf 155, react with the native protein and inhibit merozoite invasion. *Proc Natl Acad Sci U S A*, 83, 1065-9.
- BESTEIRO, S., MICHELIN, A., PONCET, J., DUBREMETZ, J. F. & LEBRUN, M. 2009. Export of a Toxoplasma gondii rhoptry neck protein complex at the host cell membrane to form the moving junction during invasion. *PLoS Pathog*, 5, e1000309.
- BISWAS, S., CHOUDHARY, P., ELIAS, S. C., MIURA, K., MILNE, K. H., DE CASSAN, S. C., COLLINS, K. A., HALSTEAD, F. D., BLISS, C. M., EWER, K. J., OSIER, F. H., HODGSON, S. H., DUNCAN, C. J., O'HARA, G. A., LONG, C. A., HILL, A. V. & DRAPER, S. J. 2014. Assessment of humoral immune responses to blood-stage malaria antigens following ChAd63-MVA immunization, controlled human malaria infection and natural exposure. *PLoS One*, 9, e107903.
- BLACK, C. G., WANG, L., WU, T. & COPPEL, R. L. 2003. Apical location of a novel EGF-like domain-containing protein of Plasmodium falciparum. *Mol Biochem Parasitol*, 127, 59-68.
- BLACK, S., SHINEFIELD, H., FIREMAN, B., LEWIS, E., RAY, P., HANSEN, J. R., ELVIN, L., ENSOR, K. M., HACKELL, J., SIBER, G., MALINOSKI, F., MADORE, D., CHANG, I., KOHBERGER, R., WATSON, W., AUSTRIAN, R. & EDWARDS, K. 2000. Efficacy, safety and immunogenicity of heptavalent pneumococcal conjugate vaccine in children. Northern California Kaiser Permanente Vaccine Study Center Group. *Pediatr Infect Dis J*, 19, 187-95.
- BLACKMAN, M. J. 2000. Proteases involved in erythrocyte invasion by the malaria parasite: function and potential as chemotherapeutic targets. *Curr Drug Targets*, 1, 59-83.
- BLACKMAN, M. J., HEIDRICH, H. G., DONACHIE, S., MCBRIDE, J. S. & HOLDER, A. A. 1990. A single fragment of a malaria merozoite surface protein remains on the parasite during red cell invasion and is the target of invasion-inhibiting antibodies. *J Exp Med*, 172, 379-82.
- BLACKMAN, M. J. & HOLDER, A. A. 1992. Secondary processing of the Plasmodium falciparum merozoite surface protein-1 (MSP1) by a calcium-dependent membrane-bound serine

- protease: shedding of MSP133 as a noncovalently associated complex with other fragments of the MSP1. *Mol Biochem Parasitol*, 50, 307-15.
- BLACKMAN, M. J., WHITTLE, H. & HOLDER, A. A. 1991. Processing of the Plasmodium falciparum major merozoite surface protein-1: identification of a 33-kilodalton secondary processing product which is shed prior to erythrocyte invasion. *Mol Biochem Parasitol*, 49, 35-44.
- BLINK, E. J., LIGHT, A., KALLIES, A., NUTT, S. L., HODGKIN, P. D. & TARLINTON, D. M. 2005. Early appearance of germinal center-derived memory B cells and plasma cells in blood after primary immunization. *J Exp Med*, 201, 545-54.
- BOJANG, K. A., OBARO, S. K., LEACH, A., D'ALESSANDRO, U., BENNETT, S., METZGER, W., BALLOU, W. R., TARGETT, G. A. & GREENWOOD, B. M. 1997. Follow-up of Gambian children recruited to a pilot safety and immunogenicity study of the malaria vaccine SPf66. *Parasite Immunol*, 19, 579-81.
- BORRE, M. B., DZIEGIEL, M., HOGH, B., PETERSEN, E., RIENECK, K., RILEY, E., MEIS, J. F., AIKAWA, M., NAKAMURA, K., HARADA, M. & ET AL. 1991. Primary structure and localization of a conserved immunogenic Plasmodium falciparum glutamate rich protein (GLURP) expressed in both the preerythrocytic and erythrocytic stages of the vertebrate life cycle. *Mol Biochem Parasitol*, 49, 119-31.
- BOUHAROUN-TAYOUN, H., ATTANATH, P., SABCHAREON, A., CHONGSUPHAJAISIDDHI, T. & DRUILHE, P. 1990. Antibodies that protect humans against Plasmodium falciparum blood stages do not on their own inhibit parasite growth and invasion in vitro, but act in cooperation with monocytes. *J Exp Med*, 172, 1633-41.
- BOUHAROUN-TAYOUN, H., OEUVRAY, C., LUNEL, F. & DRUILHE, P. 1995. Mechanisms underlying the monocyte-mediated antibody-dependent killing of Plasmodium falciparum asexual blood stages. *J Exp Med*, 182, 409-18.
- BOURGON, R., DELORENZI, M., SARGEANT, T., HODDER, A. N., CRABB, B. S. & SPEED, T. P. 2004. The serine repeat antigen (SERA) gene family phylogeny in Plasmodium: the impact of GC content and reconciliation of gene and species trees. *Mol Biol Evol*, 21, 2161-71.
- BOUSEMA, T., KREUELS, B. & GOSLING, R. 2011. Adjusting for heterogeneity of malaria transmission in longitudinal studies. *J Infect Dis*, 204, 1-3.
- BOWYER, P. W., STEWART, L. B., ASPELING-JONES, H., MENSAH-BROWN, H. E., AHOUIDI, A. D., AMAMBUA-NGWA, A., AWANDARE, G. A. & CONWAY, D. J. 2015. Variation in Plasmodium falciparum erythrocyte invasion phenotypes and merozoite ligand gene expression across different populations in areas of malaria endemicity. *Infect Immun*, 83, 2575-82.
- BOYLE, M. J., REILING, L., FENG, G., LANGER, C., OSIER, F. H., ASPELING-JONES, H., CHENG, Y. S., STUBBS, J., TETTEH, K. K., CONWAY, D. J., MCCARTHY, J. S., MULLER, I., MARSH, K., ANDERS, R. F. & BEESON, J. G. 2015. Human antibodies fix complement to inhibit Plasmodium falciparum invasion of erythrocytes and are associated with protection against malaria. *Immunity*, 42, 580-90.
- BOYLE, M. J., RICHARDS, J. S., GILSON, P. R., CHAI, W. & BEESON, J. G. 2010. Interactions with heparin-like molecules during erythrocyte invasion by Plasmodium falciparum merozoites. *Blood*, 115, 4559-68.
- BRANCH, O., CASAPIA, W. M., GAMBOA, D. V., HERNANDEZ, J. N., ALAVA, F. F., RONCAL, N., ALVAREZ, E., PEREZ, E. J. & GOTUZZO, E. 2005. Clustered local transmission and asymptomatic Plasmodium falciparum and Plasmodium vivax malaria infections in a recently emerged, hypoendemic Peruvian Amazon community. *Malar J*, 4, 27.
- BRANCH, O. H., TAKALA, S., KARIUKI, S., NAHLEN, B. L., KOLCZAK, M., HAWLEY, W. & LAL, A. A. 2001. Plasmodium falciparum genotypes, low complexity of infection, and resistance to subsequent malaria in participants in the Asembo Bay Cohort Project. *Infect Immun*, 69, 7783-92.
- BROWN, G. V., ANDERS, R. F. & KNOWLES, G. 1983. Differential effect of immunoglobulin on the in vitro growth of several isolates of Plasmodium falciparum. *Infect Immun*, 39, 1228-35.

- BROWN, G. V., ANDERS, R. F., STACE, J. D., ALPERS, M. P. & MITCHELL, G. F. 1981. Immunoprecipitation of biosynthetically-labelled proteins from different Papua New Guinea *Plasmodium falciparum* isolates by sera from individuals in the endemic area. *Parasite Immunol*, 3, 283-98.
- BROWN, G. V., CULVENOR, J. G., CREWETHER, P. E., BIANCO, A. E., COPPEL, R. L., SAINT, R. B., STAHL, H. D., KEMP, D. J. & ANDERS, R. F. 1985. Localization of the ring-infected erythrocyte surface antigen (RESA) of *Plasmodium falciparum* in merozoites and ring-infected erythrocytes. *J Exp Med*, 162, 774-9.
- BRUCE-CHWATT, L. J. 1963. A LONGITUDINAL SURVEY OF NATURAL MALARIA INFECTION IN A GROUP OF WEST AFRICAN ADULTS. *West Afr Med J*, 12, 199-217.
- BULL, P. C. & ABDI, A. I. 2016. The role of PfEMP1 as targets of naturally acquired immunity to childhood malaria: prospects for a vaccine. *Parasitology*, 143, 171-86.
- BULL, P. C., LOWE, B. S., KORTOK, M., MOLYNEUX, C. S., NEWBOLD, C. I. & MARSH, K. 1998. Parasite antigens on the infected red cell surface are targets for naturally acquired immunity to malaria. *Nat Med*, 4, 358-60.
- BUUS, S., ROCKBERG, J., FORSSTROM, B., NILSSON, P., UHLEN, M. & SCHAFER-NIELSEN, C. 2012. High-resolution mapping of linear antibody epitopes using ultrahigh-density peptide microarrays. *Mol Cell Proteomics*, 11, 1790-800.
- CAPPADORO, M., GIRIBALDI, G., O'BRIEN, E., TURRINI, F., MANNU, F., ULLIERS, D., SIMULA, G., LUZZATTO, L. & ARESE, P. 1998. Early phagocytosis of glucose-6-phosphate dehydrogenase (G6PD)-deficient erythrocytes parasitized by *Plasmodium falciparum* may explain malaria protection in G6PD deficiency. *Blood*, 92, 2527-34.
- CAPPAI, R., VAN SCHRAVENDIJK, M. R., ANDERS, R. F., PETERSON, M. G., THOMAS, L. M., COWMAN, A. F. & KEMP, D. J. 1989. Expression of the RESA gene in *Plasmodium falciparum* isolate FCR3 is prevented by a subtelomeric deletion. *Mol Cell Biol*, 9, 3584-7.
- CARLSON, J., HELMBY, H., HILL, A. V., BREWSTER, D., GREENWOOD, B. M. & WAHLGREN, M. 1990. Human cerebral malaria: association with erythrocyte rosetting and lack of anti-rosetting antibodies. *Lancet*, 336, 1457-60.
- CARLSON, J., NASH, G. B., GABUTTI, V., AL-YAMAN, F. & WAHLGREN, M. 1994. Natural protection against severe *Plasmodium falciparum* malaria due to impaired rosette formation. *Blood*, 84, 3909-14.
- CARNEVALE, E. P., KOURI, D., DARE, J. T., MCNAMARA, D. T., MUELLER, I. & ZIMMERMAN, P. A. 2007. A multiplex ligase detection reaction-fluorescent microsphere assay for simultaneous detection of single nucleotide polymorphisms associated with *Plasmodium falciparum* drug resistance. *J Clin Microbiol*, 45, 752-61.
- CAVANAGH, D. R., DODOO, D., HVIID, L., KURTZHALS, J. A., THEANDER, T. G., AKANMORI, B. D., POLLEY, S., CONWAY, D. J., KORAM, K. & MCBRIDE, J. S. 2004. Antibodies to the N-terminal block 2 of *Plasmodium falciparum* merozoite surface protein 1 are associated with protection against clinical malaria. *Infect Immun*, 72, 6492-502.
- CAVANAGH, D. R., ELHASSAN, I. M., ROPER, C., ROBINSON, V. J., GIHA, H., HOLDER, A. A., HVIID, L., THEANDER, T. G., ARNOT, D. E. & MCBRIDE, J. S. 1998. A longitudinal study of type-specific antibody responses to *Plasmodium falciparum* merozoite surface protein-1 in an area of unstable malaria in Sudan. *J Immunol*, 161, 347-59.
- CAVANAGH, D. R., KOCKEN, C. H., WHITE, J. H., COWAN, G. J., SAMUEL, K., DUBBELD, M. A., VOORBERG-VAN DER WEL, A., THOMAS, A. W., MCBRIDE, J. S. & ARNOT, D. E. 2014. Antibody responses to a novel *Plasmodium falciparum* merozoite surface protein vaccine correlate with protection against experimental malaria infection in Aotus monkeys. *PLoS One*, 9, e83704.
- CAVANAGH, D. R. & MCBRIDE, J. S. 1997. Antigenicity of recombinant proteins derived from *Plasmodium falciparum* merozoite surface protein 1. *Molecular and Biochemical Parasitology*, 85, 197-211.

- CELADA, A., CRUCHAUD, A. & PERRIN, L. H. 1982. Opsonic activity of human immune serum on in vitro phagocytosis of *Plasmodium falciparum* infected red blood cells by monocytes. *Clin Exp Immunol*, 47, 635-44.
- CELADA, A., CRUCHAUD, A. & PERRIN, L. H. 1983. Phagocytosis of *Plasmodium falciparum*-parasitized erythrocytes by human polymorphonuclear leukocytes. *J Parasitol*, 69, 49-53.
- CERTA, U., ROTMANN, D., MATILE, H. & REBER-LISKE, R. 1987. A naturally occurring gene encoding the major surface antigen precursor p190 of *Plasmodium falciparum* lacks tripeptide repeats. *EMBO J*, 6, 4137-42.
- CHANG, S. P., KRAMER, K. J., YAMAGA, K. M., KATO, A., CASE, S. E. & SIDDIQUI, W. A. 1988. *Plasmodium falciparum*: gene structure and hydropathy profile of the major merozoite surface antigen (gp195) of the Uganda-Palo Alto isolate. *Exp Parasitol*, 67, 1-11.
- CHAPPEL, J. A. & HOLDER, A. A. 1993. Monoclonal antibodies that inhibit *Plasmodium falciparum* invasion in vitro recognise the first growth factor-like domain of merozoite surface protein-1. *Mol Biochem Parasitol*, 60, 303-11.
- CHEN, L., LOPATICKI, S., RIGLAR, D. T., DEKIWADIA, C., UBOLDI, A. D., THAM, W. H., O'NEILL, M. T., RICHARD, D., BAUM, J., RALPH, S. A. & COWMAN, A. F. 2011. An EGF-like protein forms a complex with PfRh5 and is required for invasion of human erythrocytes by *Plasmodium falciparum*. *PLoS Pathog*, 7, e1002199.
- CHEN, Q., BARRAGAN, A., FERNANDEZ, V., SUNDSTROM, A., SCHLICHTERLE, M., SAHLEN, A., CARLSON, J., DATTA, S. & WAHLGREN, M. 1998. Identification of *Plasmodium falciparum* erythrocyte membrane protein 1 (PfEMP1) as the rosetting ligand of the malaria parasite *P. falciparum*. *J Exp Med*, 187, 15-23.
- CHENG, X. J., HAYASAKA, H., WATANABE, K., TAO, Y. L., LIU, J. Y., TSUKAMOTO, H., HORII, T., TANABE, K. & TACHIBANA, H. 2007. Production of high-affinity human monoclonal antibody fab fragments to the 19-kilodalton C-terminal merozoite surface protein 1 of *Plasmodium falciparum*. *Infect Immun*, 75, 3614-20.
- CHITNIS, C. E., MUKHERJEE, P., MEHTA, S., YAZDANI, S. S., DHAWAN, S., SHAKRI, A. R., BHARDWAJ, R., GUPTA, P. K., HANS, D., MAZUMDAR, S., SINGH, B., KUMAR, S., PANDEY, G., PARULEKAR, V., IMBAULT, N., SHIVYOGI, P., GODBOLE, G., MOHAN, K., LEROY, O., SINGH, K. & CHAUHAN, V. S. 2015. Phase I Clinical Trial of a Recombinant Blood Stage Vaccine Candidate for *Plasmodium falciparum* Malaria Based on MSP1 and EBA175. *PLoS One*, 10, e0117820.
- CLARK, D. L., SU, S. & DAVIDSON, E. A. 1997. Saccharide anions as inhibitors of the malaria parasite. *Glycoconj J*, 14, 473-9.
- CLARK, J. T., ANAND, R., AKOGLU, T. & MCBRIDE, J. S. 1987. Identification and characterisation of proteins associated with the rhoptry organelles of *Plasmodium falciparum* merozoites. *Parasitol Res*, 73, 425-34.
- CLARK, T. D., GREENHOUSE, B., NJAMA-MEYA, D., NZARUBARA, B., MAITEKI-SEBUGUZI, C., STAEDKE, S. G., SETO, E., KAMYA, M. R., ROSENTHAL, P. J. & DORSEY, G. 2008. Factors determining the heterogeneity of malaria incidence in children in Kampala, Uganda. *J Infect Dis*, 198, 393-400.
- COHEN, S., MC, G. I. & CARRINGTON, S. 1961. Gamma-globulin and acquired immunity to human malaria. *Nature*, 192, 733-7.
- COLEY, A. M., GUPTA, A., MURPHY, V. J., BAI, T., KIM, H., FOLEY, M., ANDERS, R. F. & BATCHELOR, A. H. 2007. Structure of the malaria antigen AMA1 in complex with a growth-inhibitory antibody. *PLoS Pathog*, 3, 1308-19.
- COLEY, A. M., PARISI, K., MASCIANTONIO, R., HOECK, J., CASEY, J. L., MURPHY, V. J., HARRIS, K. S., BATCHELOR, A. H., ANDERS, R. F. & FOLEY, M. 2006. The most polymorphic residue on *Plasmodium falciparum* apical membrane antigen 1 determines binding of an invasion-inhibitory antibody. *Infect Immun*, 74, 2628-36.
- COLLINS, C. R., WITHERS-MARTINEZ, C., BENTLEY, G. A., BATCHELOR, A. H., THOMAS, A. W. & BLACKMAN, M. J. 2007. Fine mapping of an epitope recognized by an invasion-inhibitory

- monoclonal antibody on the malaria vaccine candidate apical membrane antigen 1. *J Biol Chem*, 282, 7431-41.
- COLLINS, C. R., WITHERS-MARTINEZ, C., HACKETT, F. & BLACKMAN, M. J. 2009. An inhibitory antibody blocks interactions between components of the malarial invasion machinery. *PLoS Pathog*, 5, e1000273.
- COLLINS, W. E., ANDERS, R. F., PAPPAIOANOU, M., CAMPBELL, G. H., BROWN, G. V., KEMP, D. J., COPPEL, R. L., SKINNER, J. C., ANDRYSIAK, P. M., FAVALORO, J. M. & ET AL. 1986. Immunization of Aotus monkeys with recombinant proteins of an erythrocyte surface antigen of Plasmodium falciparum. *Nature*, 323, 259-62.
- COLLINS, W. E. & JEFFERY, G. M. 1999a. A retrospective examination of secondary sporozoite- and trophozoite-induced infections with Plasmodium falciparum: development of parasitologic and clinical immunity following secondary infection. *Am J Trop Med Hyg*, 61, 20-35.
- COLLINS, W. E. & JEFFERY, G. M. 1999b. A retrospective examination of sporozoite- and trophozoite-induced infections with Plasmodium falciparum in patients previously infected with heterologous species of Plasmodium: effect on development of parasitologic and clinical immunity. *Am J Trop Med Hyg*, 61, 36-43.
- COMBE, A., GIOVANNINI, D., CARVALHO, T. G., SPATH, S., BOISSON, B., LOUSSERT, C., THIBERGE, S., LACROIX, C., GUEIRARD, P. & MENARD, R. 2009. Clonal conditional mutagenesis in malaria parasites. *Cell Host Microbe*, 5, 386-96.
- CONSORTIUM, P. K. 2015. *Pf3k pilot data release 4* [Online].
<https://www.malariagen.net/projects/pf3k>. [Accessed].
- CONWAY, D. J., CAVANAGH, D. R., TANABE, K., ROPER, C., MIKES, Z. S., SAKIHAMA, N., BOJANG, K. A., ODUOLA, A. M., KREMSNER, P. G., ARNOT, D. E., GREENWOOD, B. M. & MCBRIDE, J. S. 2000. A principal target of human immunity to malaria identified by molecular population genetic and immunological analyses. *Nat Med*, 6, 689-92.
- COOPER, J. A. 1993. Merozoite surface antigen-I of plasmodium. *Parasitology Today*, 9, 50-54.
- COPPEL, R. L., COWMAN, A. F., ANDERS, R. F., BIANCO, A. E., SAINT, R. B., LINGELBACH, K. R., KEMP, D. J. & BROWN, G. V. 1984. Immune sera recognize on erythrocytes Plasmodium falciparum antigen composed of repeated amino acid sequences. *Nature*, 310, 789-92.
- COPPEL, R. L., COWMAN, A. F., LINGELBACH, K. R., BROWN, G. V., SAINT, R. B., KEMP, D. J. & ANDERS, R. F. 1983. Isolate-specific S-antigen of Plasmodium falciparum contains a repeated sequence of eleven amino acids. *Nature*, 306, 751-6.
- CORBETT, S. J., TOMLINSON, I. M., SONNHAMMER, E. L., BUCK, D. & WINTER, G. 1997. Sequence of the human immunoglobulin diversity (D) segment locus: a systematic analysis provides no evidence for the use of DIR segments, inverted D segments, "minor" D segments or D-D recombination. *J Mol Biol*, 270, 587-97.
- COWAN, G. J., CREASEY, A. M., DHANASARNOMBUT, K., THOMAS, A. W., REMARQUE, E. J. & CAVANAGH, D. R. 2011. A malaria vaccine based on the polymorphic block 2 region of MSP-1 that elicits a broad serotype-spanning immune response. *PLoS One*, 6, e26616.
- COWMAN, A. F., BERRY, D. & BAUM, J. 2012. The cellular and molecular basis for malaria parasite invasion of the human red blood cell. *The Journal of Cell Biology*, 198, 961-971.
- COWMAN, A. F., SAINT, R. B., COPPEL, R. L., BROWN, G. V., ANDERE, R. R. & KEMP, D. J. 1985. Conserved sequences flank variable tandem repeats in two α -antigen genes of Plasmodium falciparum. *Cell*, 40, 775-783.
- CRABB, B. S., BEESON, J. G., AMINO, R., MENARD, R., WATERS, A., WINZELER, E. A., WAHLGREN, M., FIDOCK, D. A. & NWAKA, S. 2012. Perspectives: The missing pieces. *Nature*, 484, S22-3.
- CRICK, A. J., THERON, M., TIFFERT, T., LEW, V. L., CICUTA, P. & RAYNER, J. C. 2014. Quantitation of malaria parasite-erythrocyte cell-cell interactions using optical tweezers. *Biophys J*, 107, 846-53.
- CROMPTON, P. D., KAYALA, M. A., TRAORE, B., KAYENTAO, K., ONGOIBA, A., WEISS, G. E., MOLINA, D. M., BURK, C. R., WAISBERG, M., JASINSKAS, A., TAN, X., DOUMBO, S., DOUMTABE, D.,

- KONE, Y., NARUM, D. L., LIANG, X., DOUMBO, O. K., MILLER, L. H., DOOLAN, D. L., BALDI, P., FELGNER, P. L. & PIERCE, S. K. 2010a. A prospective analysis of the Ab response to *Plasmodium falciparum* before and after a malaria season by protein microarray. *Proc Natl Acad Sci U S A*, 107, 6958-63.
- CROMPTON, P. D., MIURA, K., TRAORE, B., KAYENTAO, K., ONGOIBA, A., WEISS, G., DOUMBO, S., DOUMTABE, D., KONE, Y., HUANG, C.-Y., DOUMBO, O. K., MILLER, L. H., LONG, C. A. & PIERCE, S. K. 2010b. In Vitro Growth-Inhibitory Activity and Malaria Risk in a Cohort Study in Mali. *Infection and Immunity*, 78, 737-745.
- CROSNIER, C., BUSTAMANTE, L. Y., BARTHOLDSON, S. J., BEI, A. K., THERON, M., UCHIKAWA, M., MBOUP, S., NDIR, O., KWIATKOWSKI, D. P., DURASINGH, M. T., RAYNER, J. C. & WRIGHT, G. J. 2011. Basigin is a receptor essential for erythrocyte invasion by *Plasmodium falciparum*. *Nature*, 480, 534-537.
- CROSNIER, C., WANAGURU, M., MCDADE, B., OSIER, F. H., MARSH, K., RAYNER, J. C. & WRIGHT, G. J. 2013. A library of functional recombinant cell-surface and secreted *P. falciparum* merozoite proteins. *Mol Cell Proteomics*, 12, 3976-86.
- CUCUNUBA, Z. M., GUERRA, A. P., RAHIRANT, S. J., RIVERA, J. A., CORTES, L. J. & NICHOLLS, R. S. 2008. Asymptomatic *Plasmodium* spp. infection in Tierralta, Colombia. *Mem Inst Oswaldo Cruz*, 103, 668-73.
- D'ALESSANDRO, U., LEACH, A., DRAKELEY, C. J., BENNETT, S., OLALEYE, B. O., FEGAN, G. W., JAWARA, M., LANGEROCK, P., GEORGE, M. O., TARGETT, G. A. & ET AL. 1995. Efficacy trial of malaria vaccine SPf66 in Gambian infants. *Lancet*, 346, 462-7.
- DA SILVEIRA, L. A., DORTA, M. L., KIMURA, E. A., KATZIN, A. M., KAWAMOTO, F., TANABE, K. & FERREIRA, M. U. 1999. Allelic diversity and antibody recognition of *Plasmodium falciparum* merozoite surface protein 1 during hypoendemic malaria transmission in the Brazilian amazon region. *Infect Immun*, 67, 5906-16.
- DALTON, J. P. & MULCAHY, G. 2001. Parasite vaccines — a reality? *Veterinary Parasitology*, 98, 149-167.
- DAS, S., HERTRICH, N., PERRIN, A. J., WITHERS-MARTINEZ, C., COLLINS, C. R., JONES, M. L., WATERMEYER, J. M., FOBES, E. T., MARTIN, S. R., SAIBIL, H. R., WRIGHT, G. J., TREECK, M., EPP, C. & BLACKMAN, M. J. 2015. Processing of *Plasmodium falciparum* Merozoite Surface Protein MSP1 Activates a Spectrin-Binding Function Enabling Parasite Egress from RBCs. *Cell Host Microbe*, 18, 433-44.
- DAVIES, D. R., PADLAN, E. A. & SHERIFF, S. 1990. Antibody-antigen complexes. *Annu Rev Biochem*, 59, 439-73.
- DAVIS, M. M., CALAME, K., EARLY, P. W., LIVANT, D. L., JOHO, R., WEISSMAN, I. L. & HOOD, L. 1980. An immunoglobulin heavy-chain gene is formed by at least two recombinational events. *Nature*, 283, 733-9.
- DELPLACE, P., FORTIER, B., TRONCHIN, G., DUBREMETZ, J. F. & VERNES, A. 1987. Localization, biosynthesis, processing and isolation of a major 126 kDa antigen of the parasitophorous vacuole of *Plasmodium falciparum*. *Mol Biochem Parasitol*, 23, 193-201.
- DENT, A. E., BERGMANN-LEITNER, E. S., WILSON, D. W., TISCH, D. J., KIMMEL, R., VULULE, J., SUMBA, P. O., BEESON, J. G., ANGOV, E., MOORMANN, A. M. & KAZURA, J. W. 2008. Antibody-mediated growth inhibition of *Plasmodium falciparum*: relationship to age and protection from parasitemia in Kenyan children and adults. *PLoS One*, 3, e3557.
- DEWITT, W. S., LINDAU, P., SNYDER, T. M., SHERWOOD, A. M., VIGNALI, M., CARLSON, C. S., GREENBERG, P. D., DUERKOPP, N., EMERSON, R. O. & ROBINS, H. S. 2016. A Public Database of Memory and Naive B-Cell Receptor Sequences. *PLoS One*, 11, e0160853.
- DICKO, A., SAGARA, I., ELLIS, R. D., MIURA, K., GUINDO, O., KAMATE, B., SOGOBA, M., NIAMBELE, M. B., SISSOKO, M., BABY, M., DOLO, A., MULLEN, G. E., FAY, M. P., PIERCE, M., DIALLO, D. A., SAUL, A., MILLER, L. H. & DOUMBO, O. K. 2008. Phase 1 study of a combination AMA1 blood stage malaria vaccine in Malian children. *PLoS One*, 3, e1563.

- DISIS, M. L., DELA ROSA, C., GOODELL, V., KUAN, L. Y., CHANG, J. C., KUUS-REICHEL, K., CLAY, T. M., KIM LYERLY, H., BHATIA, S., GHANEKAR, S. A., MAINO, V. C. & MAECKER, H. T. 2006. Maximizing the retention of antigen specific lymphocyte function after cryopreservation. *J Immunol Methods*, 308, 13-8.
- DODEV, T. S., KARAGIANNIS, P., GILBERT, A. E., JOSEPHS, D. H., BOWEN, H., JAMES, L. K., BAX, H. J., BEAVIL, R., PANG, M. O., GOULD, H. J., KARAGIANNIS, S. N. & BEAVIL, A. J. 2014. A tool kit for rapid cloning and expression of recombinant antibodies. *Sci Rep*, 4, 5885.
- DODOO, D., AIKINS, A., KUSI, K. A., LAMPTEY, H., REMARQUE, E., MILLIGAN, P., BOSOMPRAH, S., CHILENGI, R., OSEI, Y. D., AKANMORI, B. D. & THEISEN, M. 2008. Cohort study of the association of antibody levels to AMA1, MSP119, MSP3 and GLURP with protection from clinical malaria in Ghanaian children. *Malar J*, 7, 142.
- DONATI, D., ZHANG, L. P., CHENE, A., CHEN, Q., FLICK, K., NYSTROM, M., WAHLGREN, M. & BEJARANO, M. T. 2004. Identification of a polyclonal B-cell activator in *Plasmodium falciparum*. *Infect Immun*, 72, 5412-8.
- DOUGLAS, A. D., BALDEVIANO, G. C., LUCAS, C. M., LUGO-ROMAN, L. A., CROSNIER, C., BARTHOLDSON, S. J., DIOUF, A., MIURA, K., LAMBERT, L. E., VENTOCILLA, J. A., LEIVA, K. P., MILNE, K. H., ILLINGWORTH, J. J., SPENCER, A. J., HJERRILD, K. A., ALANINE, D. G., TURNER, A. V., MOORHEAD, J. T., EDGEL, K. A., WU, Y., LONG, C. A., WRIGHT, G. J., LESCANO, A. G. & DRAPER, S. J. 2015. A PfrH5-based vaccine is efficacious against heterologous strain blood-stage *Plasmodium falciparum* infection in aotus monkeys. *Cell Host Microbe*, 17, 130-9.
- DOUGLAS, A. D., WILLIAMS, A. R., ILLINGWORTH, J. J., KAMUYU, G., BISWAS, S., GOODMAN, A. L., WYLLIE, D. H., CROSNIER, C., MIURA, K., WRIGHT, G. J., LONG, C. A., OSIER, F. H., MARSH, K., TURNER, A. V., HILL, A. V. & DRAPER, S. J. 2011. The blood-stage malaria antigen PfrH5 is susceptible to vaccine-inducible cross-strain neutralizing antibody. *Nat Commun*, 2, 601.
- DOUGLAS, A. D., WILLIAMS, A. R., KNUEPFER, E., ILLINGWORTH, J. J., FURZE, J. M., CROSNIER, C., CHOUDHARY, P., BUSTAMANTE, L. Y., ZAKUTANSKY, S. E., AWUAH, D. K., ALANINE, D. G., THERON, M., WORTH, A., SHIMKETS, R., RAYNER, J. C., HOLDER, A. A., WRIGHT, G. J. & DRAPER, S. J. 2014. Neutralization of *Plasmodium falciparum* merozoites by antibodies against PfrH5. *J Immunol*, 192, 245-58.
- DRAKELEY, C. J., CORRAN, P. H., COLEMAN, P. G., TONGREN, J. E., MCDONALD, S. L., CARNEIRO, I., MALIMA, R., LUSINGU, J., MANJURANO, A., NKYA, W. M., LEMNGE, M. M., COX, J., REYBURN, H. & RILEY, E. M. 2005. Estimating medium- and long-term trends in malaria transmission by using serological markers of malaria exposure. *Proc Natl Acad Sci U S A*, 102, 5108-13.
- DREW, D. R., O'DONNELL, R. A., SMITH, B. J. & CRABB, B. S. 2004. A common cross-species function for the double epidermal growth factor-like modules of the highly divergent plasmodium surface proteins MSP-1 and MSP-8. *J Biol Chem*, 279, 20147-53.
- DREYER, A. M., MATILE, H., PAPASTOGIANNIDIS, P., KAMBER, J., FAVUZZA, P., VOSS, T. S., WITTLIN, S. & PLUSCHKE, G. 2012. Passive immunoprotection of *Plasmodium falciparum*-infected mice designates the CyRPA as candidate malaria vaccine antigen. *J Immunol*, 188, 6225-37.
- DRUILHE, P. & KHUSMITH, S. 1987. Epidemiological correlation between levels of antibodies promoting merozoite phagocytosis of *Plasmodium falciparum* and malaria-immune status. *Infect Immun*, 55, 888-91.
- DRUILHE, P. & PERIGNON, J. L. 1997. A hypothesis about the chronicity of malaria infection. *Parasitol Today*, 13, 353-7.
- DRUILHE, P., SPERTINI, F., SOESOE, D., CORRADIN, G., MEJIA, P., SINGH, S., AUDRAN, R., BOUZIDI, A., OEUVRAY, C. & ROUSSILHON, C. 2005. A malaria vaccine that elicits in humans antibodies able to kill *Plasmodium falciparum*. *PLoS Med*, 2, e344.
- DUAN, J., MU, J., THERA, M. A., JOY, D., KOSAKOVSKY POND, S. L., DIEMERT, D., LONG, C., ZHOU, H., MIURA, K., OUATTARA, A., DOLO, A., DOUMBO, O., SU, X. Z. & MILLER, L. 2008. Population structure of the genes encoding the polymorphic *Plasmodium falciparum* apical membrane antigen 1: implications for vaccine design. *Proc Natl Acad Sci U S A*, 105, 7857-62.

- DUBOIS, B., DELORON, P., ASTAGNEAU, P., CHOUGNET, C. & LEPERS, J. P. 1993. Isotypic analysis of Plasmodium falciparum-specific antibodies and their relation to protection in Madagascar. *Infect Immun*, 61, 4498-500.
- DUFFY, C. W., BA, H., ASSEFA, S., AHOUIDI, A. D., DEH, Y. B., TANDIA, A., KIRSEBOM, F. C., KWIATKOWSKI, D. P. & CONWAY, D. J. 2017. Population genetic structure and adaptation of malaria parasites on the edge of endemic distribution. *Mol Ecol*.
- DUNN-WALTERS, D. K. & SPENCER, J. 1998. Strong intrinsic biases towards mutation and conservation of bases in human IgVH genes during somatic hypermutation prevent statistical analysis of antigen selection. *Immunology*, 95, 339-45.
- DURASINGH, M. T., TRIGLIA, T., RALPH, S. A., RAYNER, J. C., BARNWELL, J. W., MCFADDEN, G. I. & COWMAN, A. F. 2003. Phenotypic variation of Plasmodium falciparum merozoite proteins directs receptor targeting for invasion of human erythrocytes. *The EMBO Journal*, 22, 1047-1057.
- DUTTA, S., HAYNES, J. D., BARBOSA, A., WARE, L. A., SNAVELY, J. D., MOCH, J. K., THOMAS, A. W. & LANAR, D. E. 2005. Mode of action of invasion-inhibitory antibodies directed against apical membrane antigen 1 of Plasmodium falciparum. *Infect Immun*, 73, 2116-22.
- DUTTA, S., HAYNES, J. D., MOCH, J. K., BARBOSA, A. & LANAR, D. E. 2003. Invasion-inhibitory antibodies inhibit proteolytic processing of apical membrane antigen 1 of Plasmodium falciparum merozoites. *Proc Natl Acad Sci U S A*, 100, 12295-300.
- DUTTA, S., LEE, S. Y., BATCHELOR, A. H. & LANAR, D. E. 2007. Structural basis of antigenic escape of a malaria vaccine candidate. *Proc Natl Acad Sci U S A*, 104, 12488-93.
- EDOZIEN, J. C., GILLES, H. M. & UDEZO, I. O. K. 1962. ADULT AND CORD-BLOOD GAMMA-GLOBULIN AND IMMUNITY TO MALARIA IN NIGERIANS. *The Lancet*, 280, 951-955.
- EGAN, A. F., BLACKMAN, M. J. & KASLOW, D. C. 2000. Vaccine efficacy of recombinant Plasmodium falciparum merozoite surface protein 1 in malaria-naïve, -exposed, and/or -rechallenged Aotus vociferans monkeys. *Infect Immun*, 68, 1418-27.
- EGAN, A. F., MORRIS, J., BARNISH, G., ALLEN, S., GREENWOOD, B. M., KASLOW, D. C., HOLDER, A. A. & RILEY, E. M. 1996. Clinical immunity to Plasmodium falciparum malaria is associated with serum antibodies to the 19-kDa C-terminal fragment of the merozoite surface antigen, PfMSP-1. *J Infect Dis*, 173, 765-9.
- EKALA, M. T., JOUIN, H., LEKOULOU, F., ISSIFO, S., MERCEREAU-PUIJALON, O. & NTOUMI, F. 2002. Plasmodium falciparum merozoite surface protein 1 (MSP1): genotyping and humoral responses to allele-specific variants. *Acta Trop*, 81, 33-46.
- EL SAHLY, H. M., PATEL, S. M., ATMAR, R. L., LANFORD, T. A., DUBE, T., THOMPSON, D., SIM, B. K., LONG, C. & KEITEL, W. A. 2010. Safety and immunogenicity of a recombinant nonglycosylated erythrocyte binding antigen 175 Region II malaria vaccine in healthy adults living in an area where malaria is not endemic. *Clin Vaccine Immunol*, 17, 1552-9.
- ELIAS, S. C., CHOUDHARY, P., DE CASSAN, S. C., BISWAS, S., COLLINS, K. A., HALSTEAD, F. D., BLISS, C. M., EWER, K. J., HODGSON, S. H., DUNCAN, C. J., HILL, A. V. & DRAPER, S. J. 2014. Analysis of human B-cell responses following ChAd63-MVA MSP1 and AMA1 immunization and controlled malaria infection. *Immunology*, 141, 628-44.
- ELLIOTT, S. R. & BEESON, J. G. 2008. Estimating the Burden of Global Mortality in Children Aged <5 Years by Pathogen-Specific Causes. *Clinical Infectious Diseases*, 46, 1794-1795.
- ELLIS, R. D., WU, Y., MARTIN, L. B., SHAFFER, D., MIURA, K., AEBIG, J., ORCUTT, A., RAUSCH, K., ZHU, D., MOGENSEN, A., FAY, M. P., NARUM, D. L., LONG, C., MILLER, L. & DURBIN, A. P. 2012. Phase 1 study in malaria naïve adults of BSAM2/Alhydrogel(R)+CPG 7909, a blood stage vaccine against P. falciparum malaria. *PLoS One*, 7, e46094.
- ESCALANTE, A. A., GREBERT, H. M., CHAIYAROJ, S. C., MAGRIS, M., BISWAS, S., NAHLEN, B. L. & LAL, A. A. 2001. Polymorphism in the gene encoding the apical membrane antigen-1 (AMA-1) of Plasmodium falciparum. X. Asembo Bay Cohort Project. *Mol Biochem Parasitol*, 113, 279-87.

- ESEN, M., KREMSNER, P. G., SCHLEUCHER, R., GASSLER, M., IMOUKHUEDE, E. B., IMBAULT, N., LEROY, O., JEPSEN, S., KNUDSEN, B. W., SCHUMM, M., KNOBLOCH, J., THEISEN, M. & MORDMULLER, B. 2009. Safety and immunogenicity of GMZ2 - a MSP3-GLURP fusion protein malaria vaccine candidate. *Vaccine*, 27, 6862-8.
- FABER, B. W., HELLWIG, S., HOUARD, S., HAVELANGE, N., DROSSARD, J., MERTENS, H., CROON, A., KASTILAN, R., BYRNE, R., VAN DER WERFF, N., VAN DER EIJK, M., THOMAS, A. W., KOCKEN, C. H. & REMARQUE, E. J. 2016. Production, Quality Control, Stability and Pharmacotoxicity of a Malaria Vaccine Comprising Three Highly Similar PfAMA1 Protein Molecules to Overcome Antigenic Variation. *PLoS One*, 11, e0164053.
- FAIRHEAD, M. & HOWARTH, M. 2015. Site-specific biotinylation of purified proteins using BirA. *Methods Mol Biol*, 1266, 171-84.
- FARNERT, A., AREZ, A. P., BABIKER, H. A., BECK, H. P., BENITO, A., BJORKMAN, A., BRUCE, M. C., CONWAY, D. J., DAY, K. P., HENNING, L., MERCEREAU-PUIJALON, O., RANFORD-CARTWRIGHT, L. C., RUBIO, J. M., SNOUNOU, G., WALLIKER, D., ZWETYENGA, J. & DO ROSARIO, V. E. 2001. Genotyping of Plasmodium falciparum infections by PCR: a comparative multicentre study. *Trans R Soc Trop Med Hyg*, 95, 225-32.
- FENG, Z. P., ZHANG, X., HAN, P., ARORA, N., ANDERS, R. F. & NORTON, R. S. 2006. Abundance of intrinsically unstructured proteins in P. falciparum and other apicomplexan parasite proteomes. *Mol Biochem Parasitol*, 150, 256-67.
- FENTON, B., CLARK, J. T., KHAN, C. M., ROBINSON, J. V., WALLIKER, D., RIDLEY, R., SCAIFE, J. G. & MCBRIDE, J. S. 1991. Structural and antigenic polymorphism of the 35- to 48-kilodalton merozoite surface antigen (MSA-2) of the malaria parasite Plasmodium falciparum. *Molecular and Cellular Biology*, 11, 963-971.
- FERNANDEZ, V., HOMMEL, M., CHEN, Q., HAGBLUM, P. & WAHLGREN, M. 1999. Small, clonally variant antigens expressed on the surface of the Plasmodium falciparum-infected erythrocyte are encoded by the rif gene family and are the target of human immune responses. *J Exp Med*, 190, 1393-404.
- FLUECK, C., FRANK, G., SMITH, T., JAFARSHAD, A., NEBIE, I., SIRIMA, S. B., OLUGBILE, S., ALONSO, P., TANNER, M., DRUILHE, P., FELGER, I. & CORRADIN, G. 2009. Evaluation of two long synthetic merozoite surface protein 2 peptides as malaria vaccine candidates. *Vaccine*, 27, 2653-61.
- FOLEY, M., TILLEY, L., SAWYER, W. H. & ANDERS, R. F. 1991. The ring-infected erythrocyte surface antigen of Plasmodium falciparum associates with spectrin in the erythrocyte membrane. *Mol Biochem Parasitol*, 46, 137-47.
- FOQUET, L., HERMSEN, C. C., VAN GEMERT, G. J., VAN BRAECKEL, E., WEENING, K. E., SAUERWEIN, R., MEULEMAN, P. & LEROUX-ROELS, G. 2014. Vaccine-induced monoclonal antibodies targeting circumsporozoite protein prevent Plasmodium falciparum infection. *J Clin Invest*, 124, 140-4.
- FOWKES, F. J., RICHARDS, J. S., SIMPSON, J. A. & BEESON, J. G. 2010. The relationship between anti-merozoite antibodies and incidence of Plasmodium falciparum malaria: A systematic review and meta-analysis. *PLoS Med*, 7, e1000218.
- FOX, B. A., XING-LI, P., SUZUE, K., HORII, T. & BZIK, D. J. 1997. Plasmodium falciparum: an epitope within a highly conserved region of the 47-kDa amino-terminal domain of the serine repeat antigen is a target of parasite-inhibitory antibodies. *Exp Parasitol*, 85, 121-34.
- FRANZ, B., MAY, K. F., DRANOFF, G. & WUCHERPFENNIG, K. 2011. Ex vivo characterization and isolation of rare memory B cells with antigen tetramers. *Blood*, 118, 348-357.
- FREEMAN, R. R. & HOLDER, A. A. 1983. Surface antigens of malaria merozoites. A high molecular weight precursor is processed to an 83,000 mol wt form expressed on the surface of Plasmodium falciparum merozoites. *The Journal of Experimental Medicine*, 158, 1647-1653.
- FRUH, K., DOUMBO, O., MULLER, H. M., KOITA, O., MCBRIDE, J., CRISANTI, A., TOURE, Y. & BUJARD, H. 1991. Human antibody response to the major merozoite surface antigen of Plasmodium falciparum is strain specific and short-lived. *Infect Immun*, 59, 1319-24.

- FUGIKAHA, E., FORNAZARI, P. A., PENHALBEL RDE, S., LORENZETTI, A., MAROSO, R. D., AMORAS, J. T., SARAIVA, A. S., SILVA, R. U., BONINI-DOMINGOS, C. R., MATTOS, L. C., ROSSIT, A. R., CAVASINI, C. E. & MACHADO, R. L. 2007. Molecular screening of *Plasmodium* sp. asymptomatic carriers among transfusion centers from Brazilian Amazon region. *Rev Inst Med Trop Sao Paulo*, 49, 1-4.
- GALAMO, C. D., JAFARSHAD, A., BLANC, C. & DRUILHE, P. 2009. Anti-MSP1 Block 2 Antibodies Are Effective at Parasite Killing in an Allele-Specific Manner by Monocyte-Mediated Antibody-Dependent Cellular Inhibition. *Journal of Infectious Diseases*, 199, 1151-1154.
- GENTON, B., AL-YAMAN, F., ANDERS, R., SAUL, A., BROWN, G., PYE, D., IRVING, D. O., BRIGGS, W. R., MAI, A., GINNY, M., ADIGUMA, T., RARE, L., GIDDY, A., REBER-LISKE, R., STUERCHLER, D. & ALPERS, M. P. 2000. Safety and immunogenicity of a three-component blood-stage malaria vaccine in adults living in an endemic area of Papua New Guinea. *Vaccine*, 18, 2504-11.
- GENTON, B., AL-YAMAN, F., BETUELA, I., ANDERS, R. F., SAUL, A., BAEA, K., MELLOMBO, M., TARAICA, J., BROWN, G. V., PYE, D., IRVING, D. O., FELGER, I., BECK, H. P., SMITH, T. A. & ALPERS, M. P. 2003. Safety and immunogenicity of a three-component blood-stage malaria vaccine (MSP1, MSP2, RESA) against *Plasmodium falciparum* in Papua New Guinean children. *Vaccine*, 22, 30-41.
- GENTON, B., BETUELA, I., FELGER, I., AL-YAMAN, F., ANDERS, R. F., SAUL, A., RARE, L., BAISOR, M., LORRY, K., BROWN, G. V., PYE, D., IRVING, D. O., SMITH, T. A., BECK, H. P. & ALPERS, M. P. 2002. A recombinant blood-stage malaria vaccine reduces *Plasmodium falciparum* density and exerts selective pressure on parasite populations in a phase 1-2b trial in Papua New Guinea. *J Infect Dis*, 185, 820-7.
- GENTON, B., PLUSCHKE, G., DEGEN, L., KAMMER, A. R., WESTERFELD, N., OKITSU, S. L., SCHROLLER, S., VOUNATSOU, P., MUELLER, M. M., TANNER, M. & ZURBRIGGEN, R. 2007. A randomized placebo-controlled phase Ia malaria vaccine trial of two virosome-formulated synthetic peptides in healthy adult volunteers. *PLoS One*, 2, e1018.
- GEROLD, P., SCHOFIELD, L., BLACKMAN, M. J., HOLDER, A. A. & SCHWARZ, R. T. 1996. Structural analysis of the glycosyl-phosphatidylinositol membrane anchor of the merozoite surface proteins-1 and -2 of *Plasmodium falciparum*. *Mol Biochem Parasitol*, 75, 131-43.
- GETHING, P. W., PATIL, A. P., SMITH, D. L., GUERRA, C. A., ELYAZAR, I. R., JOHNSTON, G. L., TATEM, A. J. & HAY, S. I. 2011. A new world malaria map: *Plasmodium falciparum* endemicity in 2010. *Malar J*, 10, 378.
- GHUMRA, A., SEMBLAT, J. P., ATAIDE, R., KIFUDE, C., ADAMS, Y., CLAESSENS, A., ANONG, D. N., BULL, P. C., FENNELL, C., ARMAN, M., AMAMBUA-NGWA, A., WALTHER, M., CONWAY, D. J., KASSAMBARA, L., DOUMBO, O. K., RAZA, A. & ROWE, J. A. 2012. Induction of strain-transcending antibodies against Group A PfEMP1 surface antigens from virulent malaria parasites. *PLoS Pathog*, 8, e1002665.
- GILBERGER, T. W., THOMPSON, J. K., TRIGLIA, T., GOOD, R. T., DURAISINGH, M. T. & COWMAN, A. F. 2003. A novel erythrocyte binding antigen-175 paralogue from *Plasmodium falciparum* defines a new trypsin-resistant receptor on human erythrocytes. *J Biol Chem*, 278, 14480-6.
- GILSON, P. R., NEBL, T., VUKCEVIC, D., MORITZ, R. L., SARGEANT, T., SPEED, T. P., SCHOFIELD, L. & CRABB, B. S. 2006. Identification and stoichiometry of glycosylphosphatidylinositol-anchored membrane proteins of the human malaria parasite *Plasmodium falciparum*. *Mol Cell Proteomics*, 5, 1286-99.
- GIUDICELLI, V., CHAUME, D. & LEFRANC, M. P. 2004. IMGT/V-QUEST, an integrated software program for immunoglobulin and T cell receptor V-J and V-D-J rearrangement analysis. *Nucleic Acids Res*, 32, W435-40.
- GIUDICELLI, V., DUROUX, P., GINESTOUX, C., FOLCH, G., JABADO-MICHALOUD, J., CHAUME, D. & LEFRANC, M. P. 2006. IMGT/LIGM-DB, the IMGT comprehensive database of immunoglobulin and T cell receptor nucleotide sequences. *Nucleic Acids Res*, 34, D781-4.

- GNERRE, S., MACCALLUM, I., PRZYBYLSKI, D., RIBEIRO, F. J., BURTON, J. N., WALKER, B. J., SHARPE, T., HALL, G., SHEA, T. P., SYKES, S., BERLIN, A. M., AIRD, D., COSTELLO, M., DAZA, R., WILLIAMS, L., NICOL, R., GNIRKE, A., NUSBAUM, C., LANDER, E. S. & JAFFE, D. B. 2011. High-quality draft assemblies of mammalian genomes from massively parallel sequence data. *Proc Natl Acad Sci U S A*, 108, 1513-8.
- GOEL, S., PALMKVIST, M., MOLL, K., JOANNIN, N., LARA, P., AKHOURI, R. R., MORADI, N., OJEMALM, K., WESTMAN, M., ANGELETTI, D., KJELLIN, H., LEHTIO, J., BLIXT, O., IDESTRÖM, L., GAHMBERG, C. G., STORRY, J. R., HULT, A. K., OLSSON, M. L., VON HEIJNE, G., NILSSON, I. & WAHLGREN, M. 2015. RIFINs are adhesins implicated in severe *Plasmodium falciparum* malaria. *Nat Med*, 21, 314-7.
- GOEL, V. K., LI, X., CHEN, H., LIU, S. C., CHISHTI, A. H. & OH, S. S. 2003. Band 3 is a host receptor binding merozoite surface protein 1 during the *Plasmodium falciparum* invasion of erythrocytes. *Proc Natl Acad Sci U S A*, 100, 5164-9.
- GOOD, M. F., KASLOW, D. C. & MILLER, L. H. 1998. Pathways and strategies for developing a malaria blood-stage vaccine. *Annual Review of Immunology*, 16, 57-87.
- GOTHOT, A., GROSDENT, J. C. & PAULUS, J. M. 1996. A strategy for multiple immunophenotyping by image cytometry: model studies using latex microbeads labeled with seven streptavidin-bound fluorochromes. *Cytometry*, 24, 214-25.
- GOURLEY, T. S., WHERRY, E. J., MASOUST, D. & AHMED, R. 2004. Generation and maintenance of immunological memory. *Semin Immunol*, 16, 323-33.
- GRAVES, P. & GELBAND, H. 2006. Vaccines for preventing malaria (SPf66). *Cochrane Database Syst Rev*, Cd005966.
- GRAY, J. C., CORRAN, P. H., MANGIA, E., GAUNT, M. W., LI, Q., TETTEH, K. K., POLLEY, S. D., CONWAY, D. J., HOLDER, A. A., BACARESE-HAMILTON, T., RILEY, E. M. & CRISANTI, A. 2007. Profiling the antibody immune response against blood stage malaria vaccine candidates. *Clin Chem*, 53, 1244-53.
- GREENSTEIN, J. L., LEARY, J., HORAN, P., KAPPLER, J. W. & MARRACK, P. 1980. Flow sorting of antigen-binding B cell subsets. *J Immunol*, 124, 1472-81.
- GRIFFIN, J. T., HOLLINGSWORTH, T. D., REYBURN, H., DRAKELEY, C. J., RILEY, E. M. & GHANI, A. C. 2015. Gradual acquisition of immunity to severe malaria with increasing exposure. *Proc Biol Sci*, 282, 20142657.
- GUEVARA PATINO, J. A., HOLDER, A. A., MCBRIDE, J. S. & BLACKMAN, M. J. 1997. Antibodies that inhibit malaria merozoite surface protein-1 processing and erythrocyte invasion are blocked by naturally acquired human antibodies. *J Exp Med*, 186, 1689-99.
- GUY, A. J., IRANI, V., MACRAILD, C. A., ANDERS, R. F., NORTON, R. S., BEESON, J. G., RICHARDS, J. S. & RAMSLAND, P. A. 2015. Insights into the Immunological Properties of Intrinsically Disordered Malaria Proteins Using Proteome Scale Predictions. *PLoS One*, 10, e0141729.
- HARRIS, K. S., ADDA, C. G., KHORE, M., DREW, D. R., VALENTINI-GATT, A., FOWKES, F. J., BEESON, J. G., DUTTA, S., ANDERS, R. F. & FOLEY, M. 2014. Use of immunodampening to overcome diversity in the malarial vaccine candidate apical membrane antigen 1. *Infect Immun*, 82, 4707-17.
- HARRIS, P. K., YEOH, S., DLUZEWSKI, A. R., O'DONNELL, R. A., WITHERS-MARTINEZ, C., HACKETT, F., BANNISTER, L. H., MITCHELL, G. H. & BLACKMAN, M. J. 2005. Molecular identification of a malaria merozoite surface sheddase. *PLoS Pathog*, 1, 241-51.
- HAYAKAWA, K., ISHII, R., YAMASAKI, K., KISHIMOTO, T. & HARDY, R. R. 1987. Isolation of high-affinity memory B cells: phycoerythrin as a probe for antigen-binding cells. *Proc Natl Acad Sci U S A*, 84, 1379-83.
- HAYTON, K., GAUR, D., LIU, A., TAKAHASHI, J., HENSCHEN, B., SINGH, S., LAMBERT, L., FURUYA, T., BOUTTENOT, R., DOLL, M., NAWAZ, F., MU, J., JIANG, L., MILLER, L. H. & WELLEMS, T. E. 2008. Erythrocyte binding protein PfrH5 polymorphisms determine species-specific pathways of *Plasmodium falciparum* invasion. *Cell Host Microbe*, 4, 40-51.

- HEALER, J., CRAWFORD, S., RALPH, S., MCFADDEN, G. & COWMAN, A. F. 2002. Independent translocation of two micronemal proteins in developing *Plasmodium falciparum* merozoites. *Infect Immun*, 70, 5751-8.
- HERMSEN, C. C., VERHAGE, D. F., TELGT, D. S., TEELEN, K., BOUSEMA, J. T., ROESTENBERG, M., BOLAD, A., BERZINS, K., CORRADIN, G., LEROY, O., THEISEN, M. & SAUERWEIN, R. W. 2007. Glutamate-rich protein (GLURP) induces antibodies that inhibit in vitro growth of *Plasmodium falciparum* in a phase 1 malaria vaccine trial. *Vaccine*, 25, 2930-40.
- HERSHBERG, U., UDUMAN, M., SHLOMCHIK, M. J. & KLEINSTEIN, S. H. 2008. Improved methods for detecting selection by mutation analysis of Ig V region sequences. *Int Immunol*, 20, 683-94.
- HILL, D. L., ERIKSSON, E. M., LI WAI SUEN, C. S., CHIU, C. Y., RYG-CORNEJO, V., ROBINSON, L. J., SIBA, P. M., MUELLER, I., HANSEN, D. S. & SCHOFIELD, L. 2013. Opsonising antibodies to *P. falciparum* merozoites associated with immunity to clinical malaria. *PLoS One*, 8, e74627.
- HODDER, A. N., CREWETHER, P. E. & ANDERS, R. F. 2001. Specificity of the protective antibody response to apical membrane antigen 1. *Infect Immun*, 69, 3286-94.
- HODDER, A. N., CREWETHER, P. E., MATTHEW, M. L., REID, G. E., MORITZ, R. L., SIMPSON, R. J. & ANDERS, R. F. 1996. The disulfide bond structure of *Plasmodium* apical membrane antigen-1. *J Biol Chem*, 271, 29446-52.
- HOGH, B., PETERSEN, E., DZIEGIEL, M., DAVID, K., HANSON, A., BORRE, M., HOLM, A., VUUST, J. & JEPSEN, S. 1992. Antibodies to a recombinant glutamate-rich *Plasmodium falciparum* protein: evidence for protection of individuals living in a holoendemic area of Liberia. *Am J Trop Med Hyg*, 46, 307-13.
- HOGH, B., THOMPSON, R., ZAKIUDDIN, I. S., BOUDIN, C. & BORRE, M. 1993. Glutamate rich *Plasmodium falciparum* antigen (GLURP). *Parassitologia*, 35 Suppl, 47-50.
- HOLDER, A. A. 2009. The carboxy-terminus of merozoite surface protein 1: structure, specific antibodies and immunity to malaria. *Parasitology*, 136, 1445-56.
- HOLDER, A. A. & FREEMAN, R. R. 1984. The three major antigens on the surface of *Plasmodium falciparum* merozoites are derived from a single high molecular weight precursor. *J Exp Med*, 160, 624-9.
- HOLDER, A. A., SANDHU, J. S., HILLMAN, Y., DAVEY, L. S., NICHOLLS, S. C., COOPER, H. & LOCKYER, M. J. 1987. Processing of the precursor to the major merozoite surface antigens of *Plasmodium falciparum*. *Parasitology*, 94 (Pt 2), 199-208.
- HORROCKS, P., PINCHES, R., CHRISTODOULOU, Z., KYES, S. A. & NEWBOLD, C. I. 2004. Variable var transition rates underlie antigenic variation in malaria. *Proc Natl Acad Sci U S A*, 101, 11129-34.
- HU, J., CHEN, Z., GU, J., WAN, M., SHEN, Q., KIENY, M. P., HE, J., LI, Z., ZHANG, Q., REED, Z. H., ZHU, Y., LI, W., CAO, Y., QU, L., CAO, Z., WANG, Q., LIU, H., PAN, X., HUANG, X., ZHANG, D., XUE, X. & PAN, W. 2008. Safety and immunogenicity of a malaria vaccine, *Plasmodium falciparum* AMA-1/MSP-1 chimeric protein formulated in montanide ISA 720 in healthy adults. *PLoS One*, 3, e1952.
- HUNT, M., GALL, A., ONG, S. H., BRENER, J., FERNS, B., GOULDER, P., NASTOULI, E., KEANE, J. A., KELLAM, P. & OTTO, T. D. 2015. IVA: accurate de novo assembly of RNA virus genomes. *Bioinformatics*, 31, 2374-6.
- IDURY, R. M. & WATERMAN, M. S. 1995. A new algorithm for DNA sequence assembly. *J Comput Biol*, 2, 291-306.
- IQBAL, Z., CACCAMO, M., TURNER, I., FLICEK, P. & MCVEAN, G. 2012. De novo assembly and genotyping of variants using colored de Bruijn graphs. *Nat Genet*, 44, 226-232.
- IRANI, V., RAMSLAND, P. A., GUY, A. J., SIBA, P. M., MUELLER, I., RICHARDS, J. S. & BEESON, J. G. 2015. Acquisition of Functional Antibodies That Block the Binding of Erythrocyte-Binding Antigen 175 and Protection Against *Plasmodium falciparum* Malaria in Children. *Clin Infect Dis*, 61, 1244-52.

- JACOB, J., KELSOE, G., RAJEWSKY, K. & WEISS, U. 1991. Intracloal generation of antibody mutants in germinal centres. *Nature*, 354, 389-92.
- JANEWAY, C. A., TRAVERS, J., PAUL, WALPORT, M. & SHLOMCHIK, M. J. 2001. *Immunobiology*, 5th edition, New York, Garland Science.
- JELINEK, T., SCHULTE, C., BEHRENS, R., GROBUSCH, M. P., COULAUD, J. P., BISOFFI, Z., MATTEELLI, A., CLERINX, J., CORACHAN, M., PUENTE, S., GJORUP, I., HARMS, G., KOLLARITSCH, H., KOTLOWSKI, A., BJORKMANN, A., DELMONT, J. P., KNOBLOCH, J., NIELSEN, L. N., CUADROS, J., HATZ, C., BERAN, J., SCHMID, M. L., SCHULZE, M., LOPEZ-VELEZ, R., FLEISCHER, K., KAPAUN, A., MCWHINNEY, P., KERN, P., ATOUGIA, J., FRY, G., DA CUNHA, S. & BOECKEN, G. 2002. Imported Falciparum malaria in Europe: sentinel surveillance data from the European network on surveillance of imported infectious diseases. *Clin Infect Dis*, 34, 572-6.
- JENNINGS, R. M., JB, D. E. S., TODD, J. E., ARMSTRONG, M., FLANAGAN, K. L., RILEY, E. M. & DOHERTY, J. F. 2006. Imported Plasmodium falciparum malaria: are patients originating from disease-endemic areas less likely to develop severe disease? A prospective, observational study. *Am J Trop Med Hyg*, 75, 1195-9.
- JOHN, C. C., O'DONNELL, R. A., SUMBA, P. O., MOORMANN, A. M., DE KONING-WARD, T. F., KING, C. L., KAZURA, J. W. & CRABB, B. S. 2004. Evidence that invasion-inhibitory antibodies specific for the 19-kDa fragment of merozoite surface protein-1 (MSP-1 19) can play a protective role against blood-stage Plasmodium falciparum infection in individuals in a malaria endemic area of Africa. *J Immunol*, 173, 666-72.
- JOHNSON, G. & WU, T. T. 2001. Kabat Database and its applications: future directions. *Nucleic Acids Res*, 29, 205-6.
- JONES, D. T. & COZZETTO, D. 2015. DISOPRED3: precise disordered region predictions with annotated protein-binding activity. *Bioinformatics*, 31, 857-63.
- JOOS, C., MARRAMA, L., POLSON, H. E., CORRE, S., DIATTA, A. M., DIOUF, B., TRAPE, J. F., TALL, A., LONGACRE, S. & PERRAUT, R. 2010. Clinical protection from falciparum malaria correlates with neutrophil respiratory bursts induced by merozoites opsonized with human serum antibodies. *PLoS One*, 5, e9871.
- JOSHI, H., VALECHA, N., VERMA, A., KAUL, A., MALLICK, P. K., SHALINI, S., PRAJAPATI, S. K., SHARMA, S. K., DEV, V., BISWAS, S., NANDA, N., MALHOTRA, M. S., SUBBARAO, S. K. & DASH, A. P. 2007. Genetic structure of Plasmodium falciparum field isolates in eastern and north-eastern India. *Malar J*, 6, 60.
- JOSLING, G. A. & LLINAS, M. 2015. Sexual development in Plasmodium parasites: knowing when it's time to commit. *Nat Rev Microbiol*, 13, 573-87.
- JOUIN, H., GARRAUD, O., LONGACRE, S., BALEUX, F., MERCEREAU-PUIJALON, O. & MILON, G. 2005. Human antibodies to the polymorphic block 2 domain of the Plasmodium falciparum merozoite surface protein 1 (MSP-1) exhibit a highly skewed, peptide-specific light chain distribution. *Immunol Cell Biol*, 83, 392-5.
- JOUIN, H., ROGIER, C., TRAPE, J. F. & MERCEREAU-PUIJALON, O. 2001. Fixed, epitope-specific, cytophilic antibody response to the polymorphic block 2 domain of the Plasmodium falciparum merozoite surface antigen MSP-1 in humans living in a malaria-endemic area. *Eur J Immunol*, 31, 539-50.
- JULIANO, J. J., PORTER, K., MWAPASA, V., SEM, R., ROGERS, W. O., ARIEY, F., WONGSRICHANALAI, C., READ, A. & MESHNICK, S. R. 2010. Exposing malaria in-host diversity and estimating population diversity by capture-recapture using massively parallel pyrosequencing. *Proc Natl Acad Sci U S A*, 107, 20138-43.
- JULIUS, M. H., MASUDA, T. & HERZENBERG, L. A. 1972. Demonstration that antigen-binding cells are precursors of antibody-producing cells after purification with a fluorescence-activated cell sorter. *Proc Natl Acad Sci U S A*, 69, 1934-8.

- KANA, I. H., ADU, B., TIENDREBEOGO, R. W., SINGH, S. K., DODOO, D. & THEISEN, M. 2017. Naturally Acquired Antibodies Target the Glutamate-Rich Protein on Intact Merozoites and Predict Protection Against Febrile Malaria. *J Infect Dis*, 215, 623-630.
- KAPELSKI, S., KLOCKENBRING, T., FISCHER, R., BARTH, S. & FENDEL, R. 2014. Assessment of the neutrophilic antibody-dependent respiratory burst (ADRB) response to *Plasmodium falciparum*. *J Leukoc Biol*, 96, 1131-42.
- KARIUKI, M. M., LI, X., YAMODO, I., CHISHTI, A. H. & OH, S. S. 2005. Two *Plasmodium falciparum* merozoite proteins binding to erythrocyte band 3 form a direct complex. *Biochem Biophys Res Commun*, 338, 1690-5.
- KATOH, K., MISAWA, K., KUMA, K. & MIYATA, T. 2002. MAFFT: a novel method for rapid multiple sequence alignment based on fast Fourier transform. *Nucleic Acids Res*, 30, 3059-66.
- KAUTH, C. W., EPP, C., BUJARD, H. & LUTZ, R. 2003. The Merozoite Surface Protein 1 Complex of Human Malaria Parasite *Plasmodium falciparum*: INTERACTIONS AND ARRANGEMENTS OF SUBUNITS. *Journal of Biological Chemistry*, 278, 22257-22264.
- KAUTH, C. W., WOHLBIER, U., KERN, M., MEKONNEN, Z., LUTZ, R., MÜCKE, N., LANGOWSKI, J. & BUJARD, H. 2006. Interactions between Merozoite Surface Proteins 1, 6, and 7 of the Malaria Parasite *Plasmodium falciparum*. *Journal of Biological Chemistry*, 281, 31517-31527.
- KAWASAKI, K., MINOSHIMA, S., NAKATO, E., SHIBUYA, K., SHINTANI, A., SCHMEITS, J. L., WANG, J. & SHIMIZU, N. 1997. One-megabase sequence analysis of the human immunoglobulin lambda gene locus. *Genome Res*, 7, 250-61.
- KEITEL, W. A., KESTER, K. E., ATMAR, R. L., WHITE, A. C., BOND, N. H., HOLLAND, C. A., KRZYCH, U., PALMER, D. R., EGAN, A., DIGGS, C., BALLOU, W. R., HALL, B. F. & KASLOW, D. 1999. Phase I trial of two recombinant vaccines containing the 19kd carboxy terminal fragment of *Plasmodium falciparum* merozoite surface protein 1 (msp-1(19)) and T helper epitopes of tetanus toxoid. *Vaccine*, 18, 531-9.
- KEMP, D. J., COPPEL, R. L., COWMAN, A. F., SAINT, R. B., BROWN, G. V. & ANDERS, R. F. 1983. Expression of *Plasmodium falciparum* blood-stage antigens in *Escherichia coli*: detection with antibodies from immune humans. *Proc Natl Acad Sci U S A*, 80, 3787-91.
- KESTER, K. E., CUMMINGS, J. F., OFORI-ANYINAM, O., OCKENHOUSE, C. F., KRZYCH, U., MORIS, P., SCHWENK, R., NIELSEN, R. A., DEBEKE, Z., PINELIS, E., JUOMPAN, L., WILLIAMS, J., DOWLER, M., STEWART, V. A., WIRTZ, R. A., DUBOIS, M. C., LIEVENS, M., COHEN, J., BALLOU, W. R. & HEPPNER, D. G., JR. 2009. Randomized, double-blind, phase 2a trial of falciparum malaria vaccines RTS,S/AS01B and RTS,S/AS02A in malaria-naïve adults: safety, efficacy, and immunologic associates of protection. *J Infect Dis*, 200, 337-46.
- KHUSMITH, S. & DRUILHE, P. 1983. Cooperation between antibodies and monocytes that inhibit in vitro proliferation of *Plasmodium falciparum*. *Infect Immun*, 41, 219-23.
- KIMBI, H. K., TETTEH, K. K., POLLEY, S. D. & CONWAY, D. J. 2004. Cross-sectional study of specific antibodies to a polymorphic *Plasmodium falciparum* antigen and of parasite antigen genotypes in school children on the slope of Mount Cameroon. *Trans R Soc Trop Med Hyg*, 98, 284-9.
- KLEINSCHMIDT, I. & SHARP, B. 2001. Patterns in age-specific malaria incidence in a population exposed to low levels of malaria transmission intensity. *Trop Med Int Health*, 6, 986-91.
- KLOCK, H. E. & LESLEY, S. A. 2009. The Polymerase Incomplete Primer Extension (PIPE) method applied to high-throughput cloning and site-directed mutagenesis. *Methods Mol Biol*, 498, 91-103.
- KNAPP, B., HUNDT, E., NAU, U. & KUPPER, H. A. 1989. Molecular cloning, genomic structure and localization in a blood stage antigen of *Plasmodium falciparum* characterized by a serine stretch. *Mol Biochem Parasitol*, 32, 73-83.
- KORAM, K. A., ADU, B., OCRAN, J., KARIKARI, Y. S., ADU-AMANKWAH, S., NTIRI, M., ABUAKU, B., DODOO, D., GYAN, B., KRONMANN, K. C. & NKURUMAH, F. 2016. Safety and Immunogenicity of EBA-175 RII-NG Malaria Vaccine Administered Intramuscularly in Semi-Immune Adults: A

- Phase 1, Double-Blinded Placebo Controlled Dosage Escalation Study. *PLoS One*, 11, e0163066.
- KOREN, S. & PHILLIPPY, A. M. 2015. One chromosome, one contig: complete microbial genomes from long-read sequencing and assembly. *Curr Opin Microbiol*, 23, 110-20.
- KOUSSIS, K., WITHERS-MARTINEZ, C., YEOH, S., CHILD, M., HACKETT, F., KNUEPFER, E., JULIANO, L., WOHLBIER, U., BUJARD, H. & BLACKMAN, M. J. 2009. A multifunctional serine protease primes the malaria parasite for red blood cell invasion.
- KREUELS, B., KOBBE, R., ADJEI, S., KREUZBERG, C., VON REDEN, C., BATER, K., KLUG, S., BUSCH, W., ADJEI, O. & MAY, J. 2008. Spatial variation of malaria incidence in young children from a geographically homogeneous area with high endemicity. *J Infect Dis*, 197, 85-93.
- KRISHNARJUNA, B., ANDREW, D., MACRAILD, C. A., MORALES, R. A., BEESON, J. G., ANDERS, R. F., RICHARDS, J. S. & NORTON, R. S. 2016. Strain-transcending immune response generated by chimeras of the malaria vaccine candidate merozoite surface protein 2. *Sci Rep*, 6, 20613.
- KULANE, A., EKRE, H. P., PERLMANN, P., ROMBO, L., WAHLGREN, M. & WAHLIN, B. 1992. Effect of different fractions of heparin on Plasmodium falciparum merozoite invasion of red blood cells in vitro. *Am J Trop Med Hyg*, 46, 589-94.
- KUMAR, S., COLLINS, W., EGAN, A., YADAVA, A., GARRAUD, O., BLACKMAN, M. J., GUEVARA PATINO, J. A., DIGGS, C. & KASLOW, D. C. 2000. Immunogenicity and efficacy in aotus monkeys of four recombinant Plasmodium falciparum vaccines in multiple adjuvant formulations based on the 19-kilodalton C terminus of merozoite surface protein 1. *Infect Immun*, 68, 2215-23.
- KUMAR, S., YADAVA, A., KEISTER, D. B., TIAN, J. H., OHL, M., PERDUE-GREENFIELD, K. A., MILLER, L. H. & KASLOW, D. C. 1995. Immunogenicity and in vivo efficacy of recombinant Plasmodium falciparum merozoite surface protein-1 in Aotus monkeys. *Mol Med*, 1, 325-32.
- KUMARATILAKE, L. M., FERRANTE, A., JAEGER, T. & MORRIS-JONES, S. D. 1997. The role of complement, antibody, and tumor necrosis factor alpha in the killing of Plasmodium falciparum by the monocytic cell line THP-1. *Infect Immun*, 65, 5342-5.
- KUPPERS, R., FISCHER, U., RAJEWSKY, K. & GAUSE, A. 1992. Immunoglobulin heavy and light chain gene sequences of a human CD5 positive immunocytoma and sequences of four novel VHIII germline genes. *Immunol Lett*, 34, 57-62.
- LAMARQUE, M., BESTEIRO, S., PAPOIN, J., ROQUES, M., VULLIEZ-LE NORMAND, B., MORLON-GUYOT, J., DUBREMETZ, J. F., FAUQUENOY, S., TOMAVO, S., FABER, B. W., KOCKEN, C. H., THOMAS, A. W., BOULANGER, M. J., BENTLEY, G. A. & LEBRUN, M. 2011. The RON2-AMA1 interaction is a critical step in moving junction-dependent invasion by apicomplexan parasites. *PLoS Pathog*, 7, e1001276.
- LAMBERT, L. H., BULLOCK, J. L., COOK, S. T., MIURA, K., GARBOCZI, D. N., DIAKITE, M., FAIRHURST, R. M., SINGH, K. & LONG, C. A. 2014. Antigen reversal identifies targets of opsonizing IgGs against pregnancy-associated malaria. *Infect Immun*, 82, 4842-53.
- LANGHORNE, J., NDUNGU, F. M., SPONAAS, A.-M. & MARSH, K. 2008. Immunity to malaria: more questions than answers. *Nat Immunol*, 9, 725-732.
- LANZILLOTTI, R. & COETZER, T. L. 2006. The 10 kDa domain of human erythrocyte protein 4.1 binds the Plasmodium falciparum EBA-181 protein. *Malar J*, 5, 100.
- LARSEN, J. E., LUND, O. & NIELSEN, M. 2006. Improved method for predicting linear B-cell epitopes. *Immunome Res*, 2, 2.
- LAURENS, M. B., KOURIBA, B., BERGMANN-LEITNER, E., ANGOV, E., COULIBALY, D., DIARRA, I., DAOU, M., NIANGALY, A., BLACKWELDER, W. C., WU, Y., COHEN, J., BALLOU, W. R., VEKEMANS, J., LANAR, D. E., DUTTA, S., DIGGS, C., SOISSON, L., HEPPNER, D. G., DOUMBO, O. K., PLOWE, C. V. & THERA, M. A. 2017. Strain-specific Plasmodium falciparum growth inhibition among Malian children immunized with a blood-stage malaria vaccine. *PLoS One*, 12, e0173294.
- LAWRENCE, G., CHENG, Q. Q., REED, C., TAYLOR, D., STOWERS, A., CLOONAN, N., RZEPCHYK, C., SMILLIE, A., ANDERSON, K., POMBO, D., ALLWORTH, A., EISEN, D., ANDERS, R. & SAUL, A.

2000. Effect of vaccination with 3 recombinant asexual-stage malaria antigens on initial growth rates of *Plasmodium falciparum* in non-immune volunteers. *Vaccine*, 18, 1925-31.
- LE ROCH, K. G., JOHNSON, J. R., FLORENS, L., ZHOU, Y., SANTROSYAN, A., GRAINGER, M., YAN, S. F., WILLIAMSON, K. C., HOLDER, A. A., CARUCCI, D. J., YATES, J. R., 3RD & WINZELER, E. A. 2004. Global analysis of transcript and protein levels across the *Plasmodium falciparum* life cycle. *Genome Res*, 14, 2308-18.
- LEFFLER, E. M., BAND, G., BUSBY, G. B. J., KIVINEN, K., LE, Q. S., CLARKE, G. M., BOJANG, K. A., CONWAY, D. J., JALLOW, M., SISAY-JOOF, F., BOUGOUMA, E. C., MANGANO, V. D., MODIANO, D., SIRIMA, S. B., ACHIDI, E., APINJOH, T. O., MARSH, K., NDILA, C. M., PESHU, N., WILLIAMS, T. N., DRAKELEY, C., MANJURANO, A., REYBURN, H., RILEY, E., KACHALA, D., MOLYNEUX, M., NYIRONGO, V., TAYLOR, T., THORNTON, N., TILLEY, L., GRIMSLEY, S., DRURY, E., STALKER, J., CORNELIUS, V., HUBBART, C., JEFFREYS, A. E., ROWLANDS, K., ROCKETT, K. A., SPENCER, C. C. A. & KWIATKOWSKI, D. P. 2017. Resistance to malaria through structural variation of red blood cell invasion receptors. *Science*.
- LEFRANC, M. P., GIUDICELLI, V., DUROUX, P., JABADO-MICHALOUD, J., FOLCH, G., AOUINTI, S., CARILLON, E., DUVERGEY, H., HOULES, A., PAYSAN-LAFOSSE, T., HADI-SALJOQI, S., SASORITH, S., LEFRANC, G. & KOSSIDA, S. 2015. IMGT(R), the international ImMunoGeneTics information system(R) 25 years on. *Nucleic Acids Res*, 43, D413-22.
- LEGGETT, R. M., RAMIREZ-GONZALEZ, R. H., VERWEIJ, W., KAWASHIMA, C. G., IQBAL, Z., JONES, J. D., CACCAMO, M. & MACLEAN, D. 2013. Identifying and classifying trait linked polymorphisms in non-reference species by walking coloured de bruijn graphs. *PLoS One*, 8, e60058.
- LEPERS, J. P., DELORON, P., FONTENILLE, D. & COULANGES, P. 1988. Reappearance of falciparum malaria in central highland plateaux of Madagascar. *Lancet*, 1, 586.
- LI, H. 2013. Aligning sequence reads, clone sequences and assembly contigs with bwa-mem. *arXiv:130.3997*.
- LI, H. & DURBIN, R. 2009. Fast and accurate short read alignment with Burrows-Wheeler transform. *Bioinformatics*, 25, 1754-60.
- LI, H., HANDSAKER, B., WYSOKER, A., FENNEL, T., RUAN, J., HOMER, N., MARTH, G., ABECASIS, G. & DURBIN, R. 2009. The Sequence Alignment/Map format and SAMtools. *Bioinformatics*, 25, 2078-9.
- LI, J., MITAMURA, T., FOX, B. A., BZIK, D. J. & HORII, T. 2002. Differential localization of processed fragments of *Plasmodium falciparum* serine repeat antigen and further processing of its N-terminal 47 kDa fragment. *Parasitol Int*, 51, 343-52.
- LI, R., ZHU, H., RUAN, J., QIAN, W., FANG, X., SHI, Z., LI, Y., LI, S., SHAN, G., KRISTIANSEN, K., LI, S., YANG, H., WANG, J. & WANG, J. 2010. De novo assembly of human genomes with massively parallel short read sequencing. *Genome Res*, 20, 265-72.
- LI, X., CHEN, H., OO, T. H., DALY, T. M., BERGMAN, L. W., LIU, S. C., CHISHTI, A. H. & OH, S. S. 2004. A co-ligand complex anchors *Plasmodium falciparum* merozoites to the erythrocyte invasion receptor band 3. *J Biol Chem*, 279, 5765-71.
- LI, Z., CHEN, Y., MU, D., YUAN, J., SHI, Y., ZHANG, H., GAN, J., LI, N., HU, X., LIU, B., YANG, B. & FAN, W. 2012. Comparison of the two major classes of assembly algorithms: overlap-layout-consensus and de-bruijn-graph. *Brief Funct Genomics*, 11, 25-37.
- LIAN, S., TU, Y., WANG, Y., CHEN, X. & WANG, L. 2016. A repetitive sequence assembler based on next-generation sequencing. *Genet Mol Res*, 15.
- LIN, C. S., UBOLDI, A. D., MARAPANA, D., CZABOTAR, P. E., EPP, C., BUJARD, H., TAYLOR, N. L., PERUGINI, M. A., HODDER, A. N. & COWMAN, A. F. 2014. The merozoite surface protein 1 complex is a platform for binding to human erythrocytes by *Plasmodium falciparum*. *J Biol Chem*, 289, 25655-69.
- LINDING, R., JENSEN, L. J., DIELLA, F., BORK, P., GIBSON, T. J. & RUSSELL, R. B. 2003. Protein disorder prediction: implications for structural proteomics. *Structure*, 11, 1453-9.

- LOBO, C. A., RODRIGUEZ, M., REID, M. & LUSTIGMAN, S. 2003. Glycophorin C is the receptor for the Plasmodium falciparum erythrocyte binding ligand PfEBP-2 (baebl). *Blood*, 101, 4628-31.
- LOGAN-KLUMPLER, F. J., DE SILVA, N., BOEHME, U., ROGERS, M. B., VELARDE, G., MCQUILLAN, J. A., CARVER, T., ASLETT, M., OLSEN, C., SUBRAMANIAN, S., PHAN, I., FARRIS, C., MITRA, S., RAMASAMY, G., WANG, H., TIVEY, A., JACKSON, A., HOUSTON, R., PARKHILL, J., HOLDEN, M., HARB, O. S., BRUNK, B. P., MYLER, P. J., ROOS, D., CARRINGTON, M., SMITH, D. F., HERTZ-FOWLER, C. & BERRIMAN, M. 2012. GeneDB--an annotation database for pathogens. *Nucleic Acids Res*, 40, D98-108.
- LOPATICKI, S., MAIER, A. G., THOMPSON, J., WILSON, D. W., THAM, W. H., TRIGLIA, T., GOUT, A., SPEED, T. P., BEESON, J. G., HEALER, J. & COWMAN, A. F. 2011. Reticulocyte and erythrocyte binding-like proteins function cooperatively in invasion of human erythrocytes by malaria parasites. *Infect Immun*, 79, 1107-17.
- LUNDQUIST, R., NIELSEN, L. K., JAFARSHAD, A., SOESOE, D., CHRISTENSEN, L. H., DRUILHE, P. & DZIEGIEL, M. H. 2006. Human recombinant antibodies against Plasmodium falciparum merozoite surface protein 3 cloned from peripheral blood leukocytes of individuals with immunity to malaria demonstrate antiparasitic properties. *Infect Immun*, 74, 3222-31.
- LUXEMBURGER, C., RICCI, F., NOSTEN, F., RAIMOND, D., BATHET, S. & WHITE, N. J. 1997. The epidemiology of severe malaria in an area of low transmission in Thailand. *Trans R Soc Trop Med Hyg*, 91, 256-62.
- LYON, J. A., ANGOV, E., FAY, M. P., SULLIVAN, J. S., GIROURD, A. S., ROBINSON, S. J., BERGMANN-LEITNER, E. S., DUNCAN, E. H., DARKO, C. A., COLLINS, W. E., LONG, C. A. & BARNWELL, J. W. 2008. Protection induced by Plasmodium falciparum MSP1(42) is strain-specific, antigen and adjuvant dependent, and correlates with antibody responses. *PLoS One*, 3, e2830.
- MACLEAN, D., JONES, J. D. & STUDHOLME, D. J. 2009. Application of 'next-generation' sequencing technologies to microbial genetics. *Nat Rev Microbiol*, 7, 287-96.
- MAIER, A. G., DURASINGH, M. T., REEDER, J. C., PATEL, S. S., KAZURA, J. W., ZIMMERMAN, P. A. & COWMAN, A. F. 2003. Plasmodium falciparum erythrocyte invasion through glycophorin C and selection for Gerbich negativity in human populations. *Nat Med*, 9, 87-92.
- MALKIN, E., HU, J., LI, Z., CHEN, Z., BI, X., REED, Z., DUBOVSKY, F., LIU, J., WANG, Q., PAN, X., CHEN, T., GIERING, B., XU, Y., KANG, X., GU, J., SHEN, Q., TUCKER, K., TIERNEY, E., PAN, W., LONG, C. & CAO, Z. 2008. A phase 1 trial of PfCP2.9: an AMA1/MSP1 chimeric recombinant protein vaccine for Plasmodium falciparum malaria. *Vaccine*, 26, 6864-73.
- MARSH, K., OTOO, L., HAYES, R. J., CARSON, D. C. & GREENWOOD, B. M. 1989. Antibodies to blood stage antigens of Plasmodium falciparum in rural Gambians and their relation to protection against infection. *Trans R Soc Trop Med Hyg*, 83, 293-303.
- MARSHALL, V. M., ZHANG, L., ANDERS, R. F. & COPPEL, R. L. 1996. Diversity of the vaccine candidate AMA-1 of Plasmodium falciparum. *Mol Biochem Parasitol*, 77, 109-13.
- MASINDE, G. L., KROGSTAD, D. J., GORDON, D. M. & DUFFY, P. E. 1998. Immunization with SPf66 and subsequent infection with homologous and heterologous Plasmodium falciparum parasites. *Am J Trop Med Hyg*, 59, 600-5.
- MASKUS, D. J., BETHKE, S., SEIDEL, M., KAPELSKI, S., ADDAI-MENSAH, O., BOES, A., EDGU, G., SPIEGEL, H., REIMANN, A., FISCHER, R., BARTH, S., KLOCKENBRING, T. & FENDEL, R. 2015. Isolation, production and characterization of fully human monoclonal antibodies directed to Plasmodium falciparum MSP10. *Malar J*, 14, 276.
- MASKUS, D. J., KROLIK, M., BETHKE, S., SPIEGEL, H., KAPELSKI, S., SEIDEL, M., ADDAI-MENSAH, O., REIMANN, A., KLOCKENBRING, T., BARTH, S., FISCHER, R. & FENDEL, R. 2016. Characterization of a novel inhibitory human monoclonal antibody directed against Plasmodium falciparum Apical Membrane Antigen 1. *Sci Rep*, 6, 39462.
- MATTEELLI, A., COLOMBINI, P., GULLETTA, M., CASTELLI, F. & CAROSI, G. 1999. Epidemiological features and case management practices of imported malaria in northern Italy 1991-1995. *Trop Med Int Health*, 4, 653-7.

- MATTILA, P. S., SCHUGK, J., WU, H. & MAKELA, O. 1995. Extensive allelic sequence variation in the J region of the human immunoglobulin heavy chain gene locus. *Eur J Immunol*, 25, 2578-82.
- MATUSCHEWSKI, K. 2017. Vaccines against malaria-still a long way to go. *Febs j*.
- MAWILI-MBOUMBA, D. P., BORRMANN, S., CAVANAGH, D. R., MCBRIDE, J. S., MATSIEGUI, P. B., MISSINOU, M. A., KREMSNER, P. G. & NTOUMI, F. 2003. Antibody responses to Plasmodium falciparum merozoite surface protein-1 and efficacy of amodiaquine in Gabonese children with P. falciparum malaria. *J Infect Dis*, 187, 1137-41.
- MAYER, D. C., COFIE, J., JIANG, L., HARTL, D. L., TRACY, E., KABAT, J., MENDOZA, L. H. & MILLER, L. H. 2009. Glycophorin B is the erythrocyte receptor of Plasmodium falciparum erythrocyte-binding ligand, EBL-1. *Proc Natl Acad Sci U S A*, 106, 5348-52.
- MCBRIDE, J. S. & HEIDRICH, H. G. 1987. Fragments of the polymorphic Mr 185,000 glycoprotein from the surface of isolated Plasmodium falciparum merozoites form an antigenic complex. *Mol Biochem Parasitol*, 23, 71-84.
- MCCALLUM, F. J., PERSSON, K. E. M., MUGYENYI, C. K., FOWKES, F. J. I., SIMPSON, J. A., RICHARDS, J. S., WILLIAMS, T. N., MARSH, K. & BEESON, J. G. 2008. Acquisition of Growth-Inhibitory Antibodies against Blood-Stage Plasmodium falciparum. *PLoS ONE*, 3, e3571.
- MCCARRA, M. B., AYODO, G., SUMBA, P. O., KAZURA, J. W., MOORMANN, A. M., NARUM, D. L. & JOHN, C. C. 2011. Antibodies to Plasmodium falciparum erythrocyte-binding antigen-175 are associated with protection from clinical malaria. *Pediatr Infect Dis J*, 30, 1037-42.
- MCCARTHY, J. S., MARJASON, J., ELLIOTT, S., FAHEY, P., BANG, G., MALKIN, E., TIERNEY, E., AKED-HURDITCH, H., ADDA, C., CROSS, N., RICHARDS, J. S., FOWKES, F. J., BOYLE, M. J., LONG, C., DRUILHE, P., BEESON, J. G. & ANDERS, R. F. 2011. A phase 1 trial of MSP2-C1, a blood-stage malaria vaccine containing 2 isoforms of MSP2 formulated with Montanide(R) ISA 720. *PLoS One*, 6, e24413.
- MCGREGOR, I. A. 1964. THE PASSIVE TRANSFER OF HUMAN MALARIAL IMMUNITY. *Am J Trop Med Hyg*, 13, Suppl 237-9.
- MCHEYZER-WILLIAMS, L. J., COOL, M. & MCHEYZER-WILLIAMS, M. G. 2000. Antigen-specific B cell memory: expression and replenishment of a novel b220(-) memory b cell compartment. *J Exp Med*, 191, 1149-66.
- MCHEYZER-WILLIAMS, L. J. & MCHEYZER-WILLIAMS, M. G. 2005. Antigen-specific memory B cell development. *Annu Rev Immunol*, 23, 487-513.
- MERALDI, V., NEBIE, I., TIONO, A. B., DIALLO, D., SANOGO, E., THEISEN, M., DRUILHE, P., CORRADIN, G., MORET, R. & SIRIMA, B. S. 2004. Natural antibody response to Plasmodium falciparum Exp-1, MSP-3 and GLURP long synthetic peptides and association with protection. *Parasite Immunol*, 26, 265-72.
- METZGER, W. G., OKENU, D. M. N., CAVANAGH, D. R., ROBINSON, J. V., BOJANG, K. A., WEISS, H. A., MCBRIDE, J. S., GREENWOOD, B. M. & CONWAY, D. J. 2003. Serum IgG3 to the Plasmodium falciparum merozoite surface protein 2 is strongly associated with a reduced prospective risk of malaria. *Parasite Immunology*, 25, 307-312.
- MEYER, M. & KIRCHER, M. 2010. Illumina sequencing library preparation for highly multiplexed target capture and sequencing. *Cold Spring Harb Protoc*, 2010, pdb.prot5448.
- MIETTINEN-BAUMANN, A., STRYCH, W., MCBRIDE, J. & HEIDRICH, H. G. 1988. A 46,000 dalton Plasmodium falciparum merozoite surface glycoprotein not related to the 185,000-195,000 dalton schizont precursor molecule: isolation and characterization. *Parasitol Res*, 74, 317-23.
- MILLER, J. R., KOREN, S. & SUTTON, G. 2010. Assembly algorithms for next-generation sequencing data. *Genomics*, 95, 315-27.
- MILLER, L. H., ACKERMAN, H. C., SU, X. Z. & WELLEMS, T. E. 2013. Malaria biology and disease pathogenesis: insights for new treatments. *Nat Med*, 19, 156-67.
- MILLER, L. H., HOWARD, R. J., CARTER, R., GOOD, M. F., NUSSENZWEIG, V. & NUSSENZWEIG, R. S. 1986. Research toward malaria vaccines. *Science*, 234, 1349-56.

- MILLER, L. H., ROBERTS, T., SHAHABUDDIN, M. & MCCUTCHAN, T. F. 1993. Analysis of sequence diversity in the *Plasmodium falciparum* merozoite surface protein-1 (MSP-1). *Molecular and Biochemical Parasitology*, 59, 1-14.
- MILLS, K. E., PEARCE, J. A., CRABB, B. S. & COWMAN, A. F. 2002. Truncation of merozoite surface protein 3 disrupts its trafficking and that of acidic-basic repeat protein to the surface of *Plasmodium falciparum* merozoites. *Mol Microbiol*, 43, 1401-11.
- MIOTTO, O., AMATO, R., ASHLEY, E. A., MACINNIS, B., ALMAGRO-GARCIA, J., AMARATUNGA, C., LIM, P., MEAD, D., OYOLA, S. O., DHORDA, M., IMWONG, M., WOODROW, C., MANSKE, M., STALKER, J., DRURY, E., CAMPINO, S., AMENGA-ETEGO, L., THANH, T. N., TRAN, H. T., RINGWALD, P., BETHELL, D., NOSTEN, F., PHYO, A. P., PUKRITTAYAKAMEE, S., CHOTIVANICH, K., CHUOR, C. M., NGUON, C., SUON, S., SRENG, S., NEWTON, P. N., MAYXAY, M., KHANTHAVONG, M., HONGVANTHONG, B., HTUT, Y., HAN, K. T., KYAW, M. P., FAIZ, M. A., FANELLO, C. I., ONYAMBOKO, M., MOKUOLU, O. A., JACOB, C. G., TAKALA-HARRISON, S., PLOWE, C. V., DAY, N. P., DONDORP, A. M., SPENCER, C. C., MCVEAN, G., FAIRHURST, R. M., WHITE, N. J. & KWIATKOWSKI, D. P. 2015. Genetic architecture of artemisinin-resistant *Plasmodium falciparum*. *Nat Genet*, 47, 226-34.
- MOBEGI, V. A., DUFFY, C. W., AMAMBUA-NGWA, A., LOUA, K. M., LAMAN, E., NWAKANMA, D. C., MACINNIS, B., ASPELING-JONES, H., MURRAY, L., CLARK, T. G., KWIATKOWSKI, D. P. & CONWAY, D. J. 2014. Genome-Wide Analysis of Selection on the Malaria Parasite *Plasmodium falciparum* in West African Populations of Differing Infection Endemicity. *Molecular Biology and Evolution*, 31, 1490-1499.
- MORBACH, H., EICHHORN, E. M., LIESE, J. G. & GIRSCHICK, H. J. 2010. Reference values for B cell subpopulations from infancy to adulthood. *Clinical and Experimental Immunology*, 162, 271-279.
- MORDMULLER, B., SZYWON, K., GREUTELAERS, B., ESEN, M., MEWONO, L., TREUT, C., MURBETH, R. E., CHILENGI, R., NOOR, R., KILAMA, W. L., IMOUKHUEDE, E. B., IMBAULT, N., LEROY, O., THEISEN, M., JEPSEN, S., MILLIGAN, P., FENDEL, R., KREMSNER, P. G. & ISSIFOU, S. 2010. Safety and immunogenicity of the malaria vaccine candidate GMZ2 in malaria-exposed, adult individuals from Lambarene, Gabon. *Vaccine*, 28, 6698-703.
- MORIMATSU, K., MORIKAWA, T., TANABE, K., BZIK, D. J. & HORII, T. 1997. Sequence diversity in the amino-terminal 47 kDa fragment of the *Plasmodium falciparum* serine repeat antigen. *Mol Biochem Parasitol*, 86, 249-54.
- MOSS, D. K., REMARQUE, E. J., FABER, B. W., CAVANAGH, D. R., ARNOT, D. E., THOMAS, A. W. & HOLDER, A. A. 2012. *Plasmodium falciparum* 19-kilodalton merozoite surface protein 1 (MSP1)-specific antibodies that interfere with parasite growth in vitro can inhibit MSP1 processing, merozoite invasion, and intracellular parasite development. *Infect Immun*, 80, 1280-7.
- MUELLENBECK, M. F., UEBERHEIDE, B., AMULIC, B., EPP, A., FENYO, D., BUSSE, C. E., ESEN, M., THEISEN, M., MORDMÜLLER, B. & WARDEMAN, H. 2013. Atypical and classical memory B cells produce *Plasmodium falciparum* neutralizing antibodies. *The Journal of Experimental Medicine*, 210, 389-399.
- MUELLER, I., SCHOEPFLIN, S., SMITH, T. A., BENTON, K. L., BRETSCHER, M. T., LIN, E., KINIBORO, B., ZIMMERMAN, P. A., SPEED, T. P., SIBA, P. & FELGER, I. 2012. Force of infection is key to understanding the epidemiology of *Plasmodium falciparum* malaria in Papua New Guinean children. *Proc Natl Acad Sci U S A*, 109, 10030-5.
- MULLEN, G. E., ELLIS, R. D., MIURA, K., MALKIN, E., NOLAN, C., HAY, M., FAY, M. P., SAUL, A., ZHU, D., RAUSCH, K., MORETZ, S., ZHOU, H., LONG, C. A., MILLER, L. H. & TREANOR, J. 2008. Phase 1 trial of AMA1-C1/Alhydrogel plus CPG 7909: an asexual blood-stage vaccine for *Plasmodium falciparum* malaria. *PLoS One*, 3, e2940.

- MULLER, C. P., KREMER, J. R., BEST, J. M., DOURADO, I., TRIKI, H. & REEF, S. 2007. Reducing global disease burden of measles and rubella: report of the WHO Steering Committee on research related to measles and rubella vaccines and vaccination, 2005. *Vaccine*, 25, 1-9.
- MURHANDARWATI, E. E., BLACK, C. G., WANG, L., WEISMAN, S., KONING-WARD, T. F., BAIRD, J. K., TJITRA, E., RICHIE, T. L., CRABB, B. S. & COPPEL, R. L. 2008. Acquisition of invasion-inhibitory antibodies specific for the 19-kDa fragment of merozoite surface protein 1 in a transmigrant population requires multiple infections. *J Infect Dis*, 198, 1212-8.
- MURPHY, K., TRAVERS, P. & WALPORT, M. 2008. *Janeway's Immunobiology*, New York, Garland Science.
- MURRAY, L., MOBEGI, V. A., DUFFY, C. W., ASSEFA, S. A., KWIATKOWSKI, D. P., LAMAN, E., LOUA, K. M. & CONWAY, D. J. 2016. Microsatellite genotyping and genome-wide single nucleotide polymorphism-based indices of *Plasmodium falciparum* diversity within clinical infections. *Malar J*, 15, 275.
- NARUM, D. L. & THOMAS, A. W. 1994. Differential localization of full-length and processed forms of PF83/AMA-1 an apical membrane antigen of *Plasmodium falciparum* merozoites. *Mol Biochem Parasitol*, 67, 59-68.
- NATALIA, M., PARISA, R., ANITA, S., SIVAPRABHA, P. & ROSELLA, C. 2009. Existing antibacterial vaccines. *Dermatologic Therapy*, 22, 129-142.
- NDUATI, E. W., NG, D. H., NDUNGU, F. M., GARDNER, P., URBAN, B. C. & LANGHORNE, J. 2010. Distinct kinetics of memory B-cell and plasma-cell responses in peripheral blood following a blood-stage *Plasmodium chabaudi* infection in mice. *PLoS One*, 5, e15007.
- NDUNGU, F. M., OLOTU, A., MWACHARO, J., NYONDA, M., APFELD, J., MRAMBA, L. K., FEGAN, G. W., BEJON, P. & MARSH, K. 2012. Memory B cells are a more reliable archive for historical antimalarial responses than plasma antibodies in no-longer exposed children. *Proceedings of the National Academy of Sciences*, 109, 8247-8252.
- NEBIE, I., DIARRA, A., OUEDRAOGO, A., SOULAMA, I., BOUGOUMA, E. C., TIONO, A. B., KONATE, A. T., CHILENGI, R., THEISEN, M., DODOO, D., REMARQUE, E., BOSOMPRAH, S., MILLIGAN, P. & SIRIMA, S. B. 2008. Humoral responses to *Plasmodium falciparum* blood-stage antigens and association with incidence of clinical malaria in children living in an area of seasonal malaria transmission in Burkina Faso, West Africa. *Infect Immun*, 76, 759-66.
- NOGARO, S. I., HAFALLA, J. C., WALTHER, B., REMARQUE, E. J., TETTEH, K. K., CONWAY, D. J., RILEY, E. M. & WALTHER, M. 2011. The breadth, but not the magnitude, of circulating memory B cell responses to *P. falciparum* increases with age/exposure in an area of low transmission. *PLoS One*, 6, e25582.
- NORANATE, N., PRUGNOLLE, F., JOUIN, H., TALL, A., MARRAMA, L., SOKHNA, C., EKALA, M. T., GUILLOTTE, M., BISCHOFF, E., BOUCHIER, C., PATARAPOTIKUL, J., OHASHI, J., TRAPE, J. F., ROGIER, C. & MERCEREAU-PUIJALON, O. 2009. Population diversity and antibody selective pressure to *Plasmodium falciparum* MSP1 block2 locus in an African malaria-endemic setting. *BMC Microbiol*, 9, 219.
- NWUBA, R. I., SODEINDE, O., ANUMUDU, C. I., OMOSUN, Y. O., ODAIBO, A. B., HOLDER, A. A. & NWAGWU, M. 2002. The human immune response to *Plasmodium falciparum* includes both antibodies that inhibit merozoite surface protein 1 secondary processing and blocking antibodies. *Infect Immun*, 70, 5328-31.
- O'DONNELL, R. A., DE KONING-WARD, T. F., BURT, R. A., BOCKARIE, M., REEDER, J. C., COWMAN, A. F. & CRABB, B. S. 2001. Antibodies against merozoite surface protein (MSP)-1(19) are a major component of the invasion-inhibitory response in individuals immune to malaria. *J Exp Med*, 193, 1403-12.
- O'DONNELL, R. A., SAUL, A., COWMAN, A. F. & CRABB, B. S. 2000. Functional conservation of the malaria vaccine antigen MSP-119 across distantly related *Plasmodium* species. *Nat Med*, 6, 91-5.

- OCHOLA, L. I., TETTEH, K. K., STEWART, L. B., RIITHO, V., MARSH, K. & CONWAY, D. J. 2010. Allele frequency-based and polymorphism-versus-divergence indices of balancing selection in a new filtered set of polymorphic genes in *Plasmodium falciparum*. *Mol Biol Evol*, 27, 2344-51.
- OCKENHOUSE, C. F., ANGOV, E., KESTER, K. E., DIGGS, C., SOISSON, L., CUMMINGS, J. F., STEWART, A. V., PALMER, D. R., MAHAJAN, B., KRZYCH, U., TORNIEPORTH, N., DELCHAMBRE, M., VANHANDENHOVE, M., OFORI-ANYINAM, O., COHEN, J., LYON, J. A. & HEPPNER, D. G. 2006. Phase I safety and immunogenicity trial of FMP1/AS02A, a *Plasmodium falciparum* MSP-1 asexual blood stage vaccine. *Vaccine*, 24, 3009-17.
- OCKENHOUSE, C. F., SUN, P. F., LANAR, D. E., WELLDE, B. T., HALL, B. T., KESTER, K., STOUTE, J. A., MAGILL, A., KRZYCH, U., FARLEY, L., WIRTZ, R. A., SADOFF, J. C., KASLOW, D. C., KUMAR, S., CHURCH, L. W., CRUTCHER, J. M., WIZEL, B., HOFFMAN, S., LALVANI, A., HILL, A. V., TINE, J. A., GUITO, K. P., DE TAISNE, C., ANDERS, R., BALLOU, W. R. & ET AL. 1998. Phase I/IIa safety, immunogenicity, and efficacy trial of NYVAC-Pf7, a pox-vectored, multiantigen, multistage vaccine candidate for *Plasmodium falciparum* malaria. *J Infect Dis*, 177, 1664-73.
- OEUVRAY, C., BOUHAROUN-TAYOUN, H., GRAS-MASSE, H., BOTTIUS, E., KAIDOH, T., AIKAWA, M., FILGUEIRA, M., TARTAR, A. & DRUILHE, P. 1994. *Merozoite surface protein-3: a malaria protein inducing antibodies that promote Plasmodium falciparum killing by cooperation with blood monocytes*.
- OGUTU, B. R., APOLLO, O. J., MCKINNEY, D., OKOTH, W., SIANGLA, J., DUBOVSKY, F., TUCKER, K., WAITUMBI, J. N., DIGGS, C., WITTES, J., MALKIN, E., LEACH, A., SOISSON, L. A., MILMAN, J. B., OTIENO, L., HOLLAND, C. A., POLHEMUS, M., REMICH, S. A., OCKENHOUSE, C. F., COHEN, J., BALLOU, W. R., MARTIN, S. K., ANGOV, E., STEWART, V. A., LYON, J. A., HEPPNER, D. G. & WITHERS, M. R. 2009. Blood stage malaria vaccine eliciting high antigen-specific antibody concentrations confers no protection to young children in Western Kenya. *PLoS One*, 4, e4708.
- OKECH, B., MUJUZI, G., OGWAL, A., SHIRAI, H., HORII, T. & EGWANG, T. G. 2006. High titers of IgG antibodies against *Plasmodium falciparum* serine repeat antigen 5 (SERA5) are associated with protection against severe malaria in Ugandan children. *Am J Trop Med Hyg*, 74, 191-7.
- OKECH, B. A., NALUNKUMA, A., OKELLO, D., PANG, X. L., SUZUE, K., LI, J., HORII, T. & EGWANG, T. G. 2001. Natural human immunoglobulin G subclass responses to *Plasmodium falciparum* serine repeat antigen in Uganda. *Am J Trop Med Hyg*, 65, 912-7.
- OKENU, D. M., RILEY, E. M., BICKLE, Q. D., AGOMO, P. U., BARBOSA, A., DAUGHERTY, J. R., LANAR, D. E. & CONWAY, D. J. 2000. Analysis of human antibodies to erythrocyte binding antigen 175 of *Plasmodium falciparum*. *Infect Immun*, 68, 5559-66.
- OLOTU, A., LUSINGU, J., LEACH, A., LIEVENS, M., VEKEMANS, J., MSHAM, S., LANG, T., GOULD, J., DUBOIS, M. C., JONGERT, E., VANSADIA, P., CARTER, T., NJUGUNA, P., AWUONDO, K. O., MALABEJA, A., ABDUL, O., GESASE, S., MTURI, N., DRAKELEY, C. J., SAVARESE, B., VILLAFANA, T., LAPIERRE, D., BALLOU, W. R., COHEN, J., LEMNGE, M. M., PESHU, N., MARSH, K., RILEY, E. M., VON SEIDLEIN, L. & BEJON, P. 2011. Efficacy of RTS,S/AS01E malaria vaccine and exploratory analysis on anti-circumsporozoite antibody titres and protection in children aged 5-17 months in Kenya and Tanzania: a randomised controlled trial. *Lancet Infect Dis*, 11, 102-9.
- OSIER, F., FENG, G., BOYLE, M., LANGER, C., ZHOU, J., RICHARDS, J., MCCALLUM, F., REILING, L., JAWOROWSKI, A., ANDERS, R., MARSH, K. & BEESON, J. 2014a. Opsonic phagocytosis of *Plasmodium falciparum* merozoites: mechanism in human immunity and a correlate of protection against malaria. *BMC Medicine*, 12, 108.
- OSIER, F. H., FEGAN, G., POLLEY, S. D., MURUNGI, L., VERRA, F., TETTEH, K. K., LOWE, B., MWANGI, T., BULL, P. C., THOMAS, A. W., CAVANAGH, D. R., MCBRIDE, J. S., LANAR, D. E., MACKINNON, M. J., CONWAY, D. J. & MARSH, K. 2008. Breadth and magnitude of antibody responses to multiple *Plasmodium falciparum* merozoite antigens are associated with protection from clinical malaria. *Infect Immun*, 76, 2240-8.

- OSIER, F. H., MACKINNON, M. J., CROSNIER, C., FEGAN, G., KAMUYU, G., WANAGURU, M., OGADA, E., MCDADE, B., RAYNER, J. C., WRIGHT, G. J. & MARSH, K. 2014b. New antigens for a multicomponent blood-stage malaria vaccine. *Sci Transl Med*, 6, 247ra102.
- OSIER, F. H., POLLEY, S. D., MWANGI, T., LOWE, B., CONWAY, D. J. & MARSH, K. 2007. Naturally acquired antibodies to polymorphic and conserved epitopes of *Plasmodium falciparum* merozoite surface protein 3. *Parasite Immunol*, 29, 387-94.
- OSIER, F. H. A., MURUNGI, L. M., FEGAN, G., TUJU, J., TETTEH, K. K., BULL, P. C., CONWAY, D. J. & MARSH, K. 2010. Allele-specific antibodies to *Plasmodium falciparum* merozoite surface protein-2 and protection against clinical malaria. *Parasite Immunology*, 32, 193-201.
- OTSYULA, N., ANGOV, E., BERGMANN-LEITNER, E., KOECH, M., KHAN, F., BENNETT, J., OTIENO, L., CUMMINGS, J., ANDAGALU, B., TOSH, D., WAITUMBI, J., RICHIE, N., SHI, M., MILLER, L., OTIENO, W., OTIENO, G. A., WARE, L., HOUSE, B., GODEAUX, O., DUBOIS, M. C., OGUTU, B., BALLOU, W. R., SOISSON, L., DIGGS, C., COHEN, J., POLHEMUS, M., HEPPNER, D. G., JR., OCKENHOUSE, C. F. & SPRING, M. D. 2013. Results from tandem Phase 1 studies evaluating the safety, reactogenicity and immunogenicity of the vaccine candidate antigen *Plasmodium falciparum* FVO merozoite surface protein-1 (MSP1(42)) administered intramuscularly with adjuvant system AS01. *Malar J*, 12, 29.
- OTTO, T. D., WILINSKI, D., ASSEFA, S., KEANE, T. M., SARRY, L. R., BOHME, U., LEMIEUX, J., BARRELL, B., PAIN, A., BERRIMAN, M., NEWBOLD, C. & LLINAS, M. 2010. New insights into the blood-stage transcriptome of *Plasmodium falciparum* using RNA-Seq. *Mol Microbiol*, 76, 12-24.
- OUATTARA, A., BARRY, A. E., DUTTA, S., REMARQUE, E. J., BEESON, J. G. & PLOWE, C. V. 2015. Designing malaria vaccines to circumvent antigen variability. *Vaccine*, 33, 7506-12.
- OUATTARA, A., TAKALA-HARRISON, S., THERA, M. A., COULIBALY, D., NIANGALY, A., SAYE, R., TOLO, Y., DUTTA, S., HEPPNER, D. G., SOISSON, L., DIGGS, C. L., VEKEMANS, J., COHEN, J., BLACKWELDER, W. C., DUBE, T., LAURENS, M. B., DOUMBO, O. K. & PLOWE, C. V. 2013. Molecular basis of allele-specific efficacy of a blood-stage malaria vaccine: vaccine development implications. *J Infect Dis*, 207, 511-9.
- OULDABDALLAHI MOUKAH, M., BA, O., BA, H., OULD KHAIRY, M. L., FAYE, O., BOGREAU, H., SIMARD, F. & BASCO, L. K. 2016. Malaria in three epidemiological strata in Mauritania. *Malar J*, 15, 204.
- OWUSU-AGYEI, S., NETTEY, O. E., ZANDOH, C., SULEMANA, A., ADDA, R., AMENGA-ETEGO, S. & MBACKE, C. 2012. Demographic patterns and trends in Central Ghana: baseline indicators from the Kintampo Health and Demographic Surveillance System. *Glob Health Action*, 5, 1-11.
- PACHEBAT, J. A., LING, I. T., GRAINGER, M., TRUCCO, C., HOWELL, S., FERNANDEZ-REYES, D., GUNARATNE, R. & HOLDER, A. A. 2001. The 22 kDa component of the protein complex on the surface of *Plasmodium falciparum* merozoites is derived from a larger precursor, merozoite surface protein 7. *Mol Biochem Parasitol*, 117, 83-9.
- PALACPAC, N. M., NTEGE, E., YEKA, A., BALIKAGALA, B., SUZUKI, N., SHIRAI, H., YAGI, M., ITO, K., FUKUSHIMA, W., HIROTA, Y., NSEREKO, C., OKADA, T., KANOI, B. N., TETSUTANI, K., ARISUE, N., ITAGAKI, S., TOUGAN, T., ISHII, K. J., UEDA, S., EGWANG, T. G. & HORII, T. 2013. Phase 1b randomized trial and follow-up study in Uganda of the blood-stage malaria vaccine candidate BK-SE36. *PLoS One*, 8, e64073.
- PATEL, S. D., AHOUIDI, A. D., BEI, A. K., DIEYE, T. N., MBOUP, S., HARRISON, S. C. & DURAISINGH, M. T. 2013. *Plasmodium falciparum* merozoite surface antigen, PfrH5, elicits detectable levels of invasion-inhibiting antibodies in humans. *J Infect Dis*, 208, 1679-87.
- PERRAUT, R., MARRAMA, L., DIOUF, B., SOKHNA, C., TALL, A., NABETH, P., TRAPE, J. F., LONGACRE, S. & MERCEREAU-PUIJALON, O. 2005. Antibodies to the conserved C-terminal domain of the *Plasmodium falciparum* merozoite surface protein 1 and to the merozoite extract and their relationship with in vitro inhibitory antibodies and protection against clinical malaria in a Senegalese village. *J Infect Dis*, 191, 264-71.

- PERRAUT, R., MERCEREAU-PUIJALON, O., DIOUF, B., TALL, A., GUILLOTTE, M., LE SCANF, C., TRAPE, J. F., SPIEGEL, A. & GARRAUD, O. 2000. Seasonal fluctuation of antibody levels to *Plasmodium falciparum* parasitized red blood cell-associated antigens in two Senegalese villages with different transmission conditions. *Am J Trop Med Hyg*, 62, 746-51.
- PERSSON, K. E., FOWKES, F. J., MCCALLUM, F. J., GICHERU, N., REILING, L., RICHARDS, J. S., WILSON, D. W., LOPATICKI, S., COWMAN, A. F., MARSH, K. & BEESON, J. G. 2013. Erythrocyte-binding antigens of *Plasmodium falciparum* are targets of human inhibitory antibodies and function to evade naturally acquired immunity. *J Immunol*, 191, 785-94.
- PERSSON, K. E., MCCALLUM, F. J., REILING, L., LISTER, N. A., STUBBS, J., COWMAN, A. F., MARSH, K. & BEESON, J. G. 2008. Variation in use of erythrocyte invasion pathways by *Plasmodium falciparum* mediates evasion of human inhibitory antibodies. *J Clin Invest*, 118, 342-51.
- PETERSEN, E., HOGH, B., MARBIAH, N. T., PERLMANN, H., WILLCOX, M., DOLOPAIE, E., HANSON, A. P., BJORKMAN, A. & PERLMANN, P. 1990. A longitudinal study of antibodies to the *Plasmodium falciparum* antigen Pf155/RESA and immunity to malaria infection in adult Liberians. *Trans R Soc Trop Med Hyg*, 84, 339-45.
- PETERSON, M. G., MARSHALL, V. M., SMYTHE, J. A., CREWETHER, P. E., LEW, A., SILVA, A., ANDERS, R. F. & KEMP, D. J. 1989. Integral membrane protein located in the apical complex of *Plasmodium falciparum*. *Mol Cell Biol*, 9, 3151-4.
- PHAM, P., BRANSTEITZER, R., PETRUSKA, J. & GOODMAN, M. F. 2003. Processive AID-catalysed cytosine deamination on single-stranded DNA simulates somatic hypermutation. *Nature*, 424, 103-7.
- PHILLIPS, R. S., TRIGG, P. I., SCOTT-FINNIGAN, T. J. & BARTHOLOMEW, R. K. 1972. Culture of *Plasmodium falciparum* in vitro: a subculture technique used for demonstrating antiparasitoid activity in serum from some Gambians, resident in an endemic malarious area. *Parasitology*, 65, 525-35.
- PHYO, A. P., NKHOMA, S., STEPNIIEWSKA, K., ASHLEY, E. A., NAIR, S., MCGREADY, R., LER MOO, C., AL-SAAI, S., DONDORP, A. M., LWIN, K. M., SINGHASIVANON, P., DAY, N. P., WHITE, N. J., ANDERSON, T. J. & NOSTEN, F. 2012. Emergence of artemisinin-resistant malaria on the western border of Thailand: a longitudinal study. *Lancet*, 379, 1960-6.
- PIZARRO, J. C., VULLIEZ-LE NORMAND, B., CHESNE-SECK, M. L., COLLINS, C. R., WITHERS-MARTINEZ, C., HACKETT, F., BLACKMAN, M. J., FABER, B. W., REMARQUE, E. J., KOCKEN, C. H., THOMAS, A. W. & BENTLEY, G. A. 2005. Crystal structure of the malaria vaccine candidate apical membrane antigen 1. *Science*, 308, 408-11.
- POLHEMUS, M. E., MAGILL, A. J., CUMMINGS, J. F., KESTER, K. E., OCKENHOUSE, C. F., LANAR, D. E., DUTTA, S., BARBOSA, A., SOISSON, L., DIGGS, C. L., ROBINSON, S. A., HAYNES, J. D., STEWART, V. A., WARE, L. A., BRANDO, C., KRZYCH, U., BOWDEN, R. A., COHEN, J. D., DUBOIS, M. C., OFORI-ANYINAM, O., DE-KOCK, E., BALLOU, W. R. & HEPPNER, D. G., JR. 2007. Phase I dose escalation safety and immunogenicity trial of *Plasmodium falciparum* apical membrane protein (AMA-1) FMP2.1, adjuvanted with AS02A, in malaria-naïve adults at the Walter Reed Army Institute of Research. *Vaccine*, 25, 4203-12.
- POLLEY, S. D., CHOKEJINDACHAI, W. & CONWAY, D. J. 2003a. Allele frequency-based analyses robustly map sequence sites under balancing selection in a malaria vaccine candidate antigen. *Genetics*, 165, 555-61.
- POLLEY, S. D. & CONWAY, D. J. 2001. Strong diversifying selection on domains of the *Plasmodium falciparum* apical membrane antigen 1 gene. *Genetics*, 158, 1505-12.
- POLLEY, S. D., CONWAY, D. J., CAVANAGH, D. R., MCBRIDE, J. S., LOWE, B. S., WILLIAMS, T. N., MWANGI, T. W. & MARSH, K. 2006. High levels of serum antibodies to merozoite surface protein 2 of *Plasmodium falciparum* are associated with reduced risk of clinical malaria in coastal Kenya. *Vaccine*, 24, 4233-4246.
- POLLEY, S. D., MWANGI, T., KOCKEN, C. H., THOMAS, A. W., DUTTA, S., LANAR, D. E., REMARQUE, E., ROSS, A., WILLIAMS, T. N., MWAMBINGU, G., LOWE, B., CONWAY, D. J. & MARSH, K. 2004.

- Human antibodies to recombinant protein constructs of Plasmodium falciparum Apical Membrane Antigen 1 (AMA1) and their associations with protection from malaria. *Vaccine*, 23, 718-28.
- POLLEY, S. D., TETTEH, K. K., CAVANAGH, D. R., PEARCE, R. J., LLOYD, J. M., BOJANG, K. A., OKENU, D. M., GREENWOOD, B. M., MCBRIDE, J. S. & CONWAY, D. J. 2003b. Repeat sequences in block 2 of Plasmodium falciparum merozoite surface protein 1 are targets of antibodies associated with protection from malaria. *Infect Immun*, 71, 1833-42.
- POLLEY, S. D., TETTEH, K. K. A., LLOYD, J. M., AKPOGHENETA, O. J., GREENWOOD, B. M., BOJANG, K. A. & CONWAY, D. J. 2007. Plasmodium falciparum Merozoite Surface Protein 3 Is a Target of Allele-Specific Immunity and Alleles Are Maintained by Natural Selection. *Journal of Infectious Diseases*, 195, 279-287.
- POMBO, D. J., LAWRENCE, G., HIRUNPETCHARAT, C., RZEPCHYK, C., BRYDEN, M., CLOONAN, N., ANDERSON, K., MAHAKUNKIJCHAROEN, Y., MARTIN, L. B., WILSON, D., ELLIOTT, S., ELLIOTT, S., EISEN, D. P., WEINBERG, J. B., SAUL, A. & GOOD, M. F. 2002. Immunity to malaria after administration of ultra-low doses of red cells infected with Plasmodium falciparum. *Lancet*, 360, 610-7.
- PROTOPOPOFF, N., MATOWO, J., MALIMA, R., KAVISHE, R., KAAYA, R., WRIGHT, A., WEST, P., KLEINSCHMIDT, I., KISINZA, W., MOSHA, F. & ROWLAND, M. 2013. High level of resistance in the mosquito Anopheles gambiae to pyrethroid insecticides and reduced susceptibility to bendiocarb in north-western Tanzania. *Malaria Journal*, 12, 149.
- PUENTES, A., GARCIA, J., OCAMPO, M., RODRIGUEZ, L., VERA, R., CURTIDOR, H., LOPEZ, R., SUAREZ, J., VALBUENA, J., VANEGAS, M., GUZMAN, F., TOVAR, D. & PATARROYO, M. E. 2003. P. falciparum: merozoite surface protein-8 peptides bind specifically to human erythrocytes. *Peptides*, 24, 1015-23.
- PUENTES, A., OCAMPO, M., RODRIGUEZ, L. E., VERA, R., VALBUENA, J., CURTIDOR, H., GARCIA, J., LOPEZ, R., TOVAR, D., CORTES, J., RIVERA, Z. & PATARROYO, M. E. 2005. Identifying Plasmodium falciparum merozoite surface protein-10 human erythrocyte specific binding regions. *Biochimie*, 87, 461-72.
- PUMPAIBOOL, T., ARNATHAU, C., DURAND, P., KANCHANAKHAN, N., SIRIPOON, N., SUEGORN, A., SITTHI-AMORN, C., RENAUD, F. & HARNYUTTANAKORN, P. 2009. Genetic diversity and population structure of Plasmodium falciparum in Thailand, a low transmission country. *Malar J*, 8, 155.
- RAVETCH, J. V., SIEBENLIST, U., KORSMEYER, S., WALDMANN, T. & LEDER, P. 1981. Structure of the human immunoglobulin mu locus: characterization of embryonic and rearranged J and D genes. *Cell*, 27, 583-91.
- RAYNER, J. C., GALINSKI, M. R., INGRAVALLO, P. & BARNWELL, J. W. 2000. Two Plasmodium falciparum genes express merozoite proteins that are related to Plasmodium vivax and Plasmodium yoelii adhesive proteins involved in host cell selection and invasion. *Proc Natl Acad Sci U S A*, 97, 9648-53.
- REDDY, K. S., AMLABU, E., PANDEY, A. K., MITRA, P., CHAUHAN, V. S. & GAUR, D. 2015. Multiprotein complex between the GPI-anchored CyRPA with PfrH5 and PfrRipr is crucial for Plasmodium falciparum erythrocyte invasion. *Proc Natl Acad Sci U S A*, 112, 1179-84.
- REESE, R. T., MOTYL, M. R. & HOFER-WARBINEK, R. 1981. Reaction of immune sera with components of the human malarial parasite, Plasmodium falciparum. *Am J Trop Med Hyg*, 30, 1168-78.
- REILING, L., RICHARDS, J. S., FOWKES, F. J., BARRY, A. E., TRIGLIA, T., CHOKEJINDACHAI, W., MICHON, P., TAVUL, L., SIBA, P. M., COWMAN, A. F., MUELLER, I. & BEESON, J. G. 2010. Evidence that the erythrocyte invasion ligand PfrH2 is a target of protective immunity against Plasmodium falciparum malaria. *J Immunol*, 185, 6157-67.
- REMARQUE, E. J., FABER, B. W., KOCKEN, C. H. & THOMAS, A. W. 2008. A diversity-covering approach to immunization with Plasmodium falciparum apical membrane antigen 1 induces

- broader allelic recognition and growth inhibition responses in rabbits. *Infect Immun*, 76, 2660-70.
- RICHARDS, J. S., ARUMUGAM, T. U., REILING, L., HEALER, J., HODDER, A. N., FOWKES, F. J., CROSS, N., LANGER, C., TAKEO, S., UBOLDI, A. D., THOMPSON, J. K., GILSON, P. R., COPPEL, R. L., SIBA, P. M., KING, C. L., TORII, M., CHITNIS, C. E., NARUM, D. L., MUELLER, I., CRABB, B. S., COWMAN, A. F., TSUBOI, T. & BEESON, J. G. 2013. Identification and prioritization of merozoite antigens as targets of protective human immunity to *Plasmodium falciparum* malaria for vaccine and biomarker development. *J Immunol*, 191, 795-809.
- RICHARDS, J. S. & BEESON, J. G. 2009. The future for blood-stage vaccines against malaria. *Immunol Cell Biol*, 87, 377-390.
- RICHARDS, J. S., STANISIC, D. I., FOWKES, F. J., TAVUL, L., DABOD, E., THOMPSON, J. K., KUMAR, S., CHITNIS, C. E., NARUM, D. L., MICHON, P., SIBA, P. M., COWMAN, A. F., MUELLER, I. & BEESON, J. G. 2010. Association between naturally acquired antibodies to erythrocyte-binding antigens of *Plasmodium falciparum* and protection from malaria and high-density parasitemia. *Clin Infect Dis*, 51, e50-60.
- RICHE, T. L. & SAUL, A. 2002. Progress and challenges for malaria vaccines. *Nature*, 415, 694-701.
- RIGLAR, D. T., RICHARD, D., WILSON, D. W., BOYLE, M. J., DEKIWADIA, C., TURNBULL, L., ANGRISANO, F., MARAPANA, D. S., ROGERS, K. L., WHITCHURCH, C. B., BEESON, J. G., COWMAN, A. F., RALPH, S. A. & BAUM, J. 2011. Super-resolution dissection of coordinated events during malaria parasite invasion of the human erythrocyte. *Cell Host Microbe*, 9, 9-20.
- RILEY, E. M. 1996. The role of MHC- and non-MHC-associated genes in determining the human immune response to malaria antigens. *Parasitology*, 112 Suppl, S39-51.
- ROBINSON, J., WALLER, M. J., PARHAM, P., BODMER, J. G. & MARSH, S. G. 2001. IMGT/HLA Database--a sequence database for the human major histocompatibility complex. *Nucleic Acids Res*, 29, 210-3.
- ROCK, E. P., SIBBALD, P. R., DAVIS, M. M. & CHIEN, Y. H. 1994. CDR3 length in antigen-specific immune receptors. *J Exp Med*, 179, 323-8.
- ROESTENBERG, M., REMARQUE, E., DE JONGE, E., HERMSEN, R., BLYTHMAN, H., LEROY, O., IMOUKHUEDE, E., JEPSEN, S., OFORI-ANYINAM, O., FABER, B., KOCKEN, C. H., ARNOLD, M., WALRAVEN, V., TELEN, K., ROEFFEN, W., DE MAST, Q., BALLOU, W. R., COHEN, J., DUBOIS, M. C., ASCARATEIL, S., VAN DER VEN, A., THOMAS, A. & SAUERWEIN, R. 2008. Safety and immunogenicity of a recombinant *Plasmodium falciparum* AMA1 malaria vaccine adjuvanted with Alhydrogel, Montanide ISA 720 or AS02. *PLoS One*, 3, e3960.
- ROLLIER, C. S., REYES-SANDOVAL, A., COTTINGHAM, M. G., EWER, K. & HILL, A. V. 2011. Viral vectors as vaccine platforms: deployment in sight. *Curr Opin Immunol*, 23, 377-82.
- ROMI, R., RAZAIARIMANGA, M. C., RAHARIMANGA, R., RAKOTONDRAIBE, E. M., RANAIVO, L. H., PIETRA, V., RAVELOSON, A. & MAJORI, G. 2002. Impact of the malaria control campaign (1993-1998) in the highlands of Madagascar: parasitological and entomological data. *Am J Trop Med Hyg*, 66, 2-6.
- ROPER, C., ELHASSAN, I. M., HVIID, L., GIHA, H., RICHARDSON, W., BABIKER, H., SATTI, G. M., THEANDER, T. G. & ARNOT, D. E. 1996. Detection of very low level *Plasmodium falciparum* infections using the nested polymerase chain reaction and a reassessment of the epidemiology of unstable malaria in Sudan. *Am J Trop Med Hyg*, 54, 325-31.
- ROPER, M. H., TORRES, R. S., GOICOCHEA, C. G., ANDERSEN, E. M., GUARDA, J. S., CALAMPA, C., HIGHTOWER, A. W. & MAGILL, A. J. 2000. The epidemiology of malaria in an epidemic area of the Peruvian Amazon. *Am J Trop Med Hyg*, 62, 247-56.
- ROSHANRAVAN, B., KARI, E., GILMAN, R. H., CABRERA, L., LEE, E., METCALFE, J., CALDERON, M., LESCANO, A. G., MONTENEGRO, S. H., CALAMPA, C. & VINETZ, J. M. 2003. Endemic malaria in the Peruvian Amazon region of Iquitos. *Am J Trop Med Hyg*, 69, 45-52.
- ROWE, A., OBEIRO, J., NEWBOLD, C. I. & MARSH, K. 1995. *Plasmodium falciparum* rosetting is associated with malaria severity in Kenya. *Infect Immun*, 63, 2323-6.

- ROWE, J. A., MOULDS, J. M., NEWBOLD, C. I. & MILLER, L. H. 1997. P. falciparum rosetting mediated by a parasite-variant erythrocyte membrane protein and complement-receptor 1. *Nature*, 388, 292-5.
- RTS, S. C. T. P. 2015. Efficacy and safety of RTS,S/AS01 malaria vaccine with or without a booster dose in infants and children in Africa: final results of a phase 3, individually randomised, controlled trial. *Lancet*, 386, 31-45.
- RUBY, J. G., BELLARE, P. & DERISI, J. L. 2013. PRICE: software for the targeted assembly of components of (Meta) genomic sequence data. *G3 (Bethesda)*, 3, 865-80.
- RUTLEDGE, G. G., BOEHME, U., SANDERS, M., REID, A. J., MAIGA-ASCOFARE, O., DJIMDE, A. A., APINJOH, T. O., AMENGA-ETEGO, L., MANSKE, M., BARNWELL, J. W., RENAUD, F., OLLOMO, B., PRUGNOLLE, F., ANSTEY, N. M., AUBURN, S., PRICE, R. N., MCCARTHY, J. S., KWIATKOWSKI, D. P., NEWBOLD, C. I., BERRIMAN, M. & OTTO, T. D. 2016. Elusive Plasmodium Species Complete the Human Malaria Genome Set. *bioRxiv*.
- SABCHAREON, A., BURNOUF, T., OUATTARA, D., ATTANATH, P., BOUHAROUN-TAYOUN, H., CHANTAVANICH, P., FOUCAULT, C., CHONGSUPHAJAISIDDHI, T. & DRUILHE, P. 1991. Parasitologic and clinical human response to immunoglobulin administration in falciparum malaria. *Am J Trop Med Hyg*, 45, 297-308.
- SAGARA, I., DICKO, A., ELLIS, R. D., FAY, M. P., DIAWARA, S. I., ASSADOU, M. H., SISSOKO, M. S., KONE, M., DIALLO, A. I., SAYE, R., GUINDO, M. A., KANTE, O., NIAMBELE, M. B., MIURA, K., MULLEN, G. E., PIERCE, M., MARTIN, L. B., DOLO, A., DIALLO, D. A., DOUMBO, O. K., MILLER, L. H. & SAUL, A. 2009. A randomized controlled phase 2 trial of the blood stage AMA1-C1/Alhydrogel malaria vaccine in children in Mali. *Vaccine*, 27, 3090-8.
- SAKAMOTO, H., TAKEO, S., MAIER, A. G., SATTABONGKOT, J., COWMAN, A. F. & TSUBOI, T. 2012. Antibodies against a Plasmodium falciparum antigen PfMSPDBL1 inhibit merozoite invasion into human erythrocytes. *Vaccine*, 30, 1972-80.
- SALK, J. E. 1953. Studies in human subjects on active immunization against poliomyelitis. I. A preliminary report of experiments in progress. *J Am Med Assoc*, 151, 1081-98.
- SALMON, D., VILDE, J. L., ANDRIEU, B., SIMONOVIC, R. & LEBRAS, J. 1986. Role of immune serum and complement in stimulation of the metabolic burst of human neutrophils by Plasmodium falciparum. *Infect Immun*, 51, 801-6.
- SANDERS, P. R., GILSON, P. R., CANTIN, G. T., GREENBAUM, D. C., NEBL, T., CARUCCI, D. J., MCCONVILLE, M. J., SCHOFIELD, L., HODDER, A. N., YATES, J. R., 3RD & CRABB, B. S. 2005. Distinct protein classes including novel merozoite surface antigens in Raft-like membranes of Plasmodium falciparum. *J Biol Chem*, 280, 40169-76.
- SANGER, F. & COULSON, A. R. 1975. A rapid method for determining sequences in DNA by primed synthesis with DNA polymerase. *J Mol Biol*, 94, 441-8.
- SANGER, F., NICKLEN, S. & COULSON, A. R. 1977. DNA sequencing with chain-terminating inhibitors. *Proc Natl Acad Sci U S A*, 74, 5463-7.
- SARR, J. B., PELLEAU, S., TOLY, C., GUITARD, J., KONATE, L., DELORON, P., GARCIA, A. & MIGOT-NABIAS, F. 2006. Impact of red blood cell polymorphisms on the antibody response to Plasmodium falciparum in Senegal. *Microbes Infect*, 8, 1260-8.
- SAUL, A., LAWRENCE, G., ALLWORTH, A., ELLIOTT, S., ANDERSON, K., RZEPCHYK, C., MARTIN, L. B., TAYLOR, D., EISEN, D. P., IRVING, D. O., PYE, D., CREWETHER, P. E., HODDER, A. N., MURPHY, V. J. & ANDERS, R. F. 2005. A human phase 1 vaccine clinical trial of the Plasmodium falciparum malaria vaccine candidate apical membrane antigen 1 in Montanide ISA720 adjuvant. *Vaccine*, 23, 3076-83.
- SAUL, A., LAWRENCE, G., SMILLIE, A., RZEPCHYK, C. M., REED, C., TAYLOR, D., ANDERSON, K., STOWERS, A., KEMP, R., ALLWORTH, A., ANDERS, R. F., BROWN, G. V., PYE, D., SCHOOF, P., IRVING, D. O., DYER, S. L., WOODROW, G. C., BRIGGS, W. R., REBER, R. & STURCHLER, D. 1999. Human phase I vaccine trials of 3 recombinant asexual stage malaria antigens with Montanide ISA720 adjuvant. *Vaccine*, 17, 3145-59.

- SCANLAN, C. N., OFFER, J., ZITZMANN, N. & DWEK, R. A. 2007. Exploiting the defensive sugars of HIV-1 for drug and vaccine design. *Nature*, 446, 1038-45.
- SCHATZ, P. J. 1993. Use of peptide libraries to map the substrate specificity of a peptide-modifying enzyme: a 13 residue consensus peptide specifies biotinylation in *Escherichia coli*. *Biotechnology (N Y)*, 11, 1138-43.
- SCHERF, A., LOPEZ-RUBIO, J. J. & RIVIERE, L. 2008. Antigenic variation in *Plasmodium falciparum*. *Annu Rev Microbiol*, 62, 445-70.
- SCHROEDER, H. W., JR. & CAVACINI, L. 2010. Structure and function of immunoglobulins. *J Allergy Clin Immunol*, 125, S41-52.
- SCOPEL, K. K., DA SILVA-NUNES, M., MALAFRONTÉ, R. S., BRAGA, E. M. & FERREIRA, M. U. 2007. Variant-specific antibodies to merozoite surface protein 2 and clinical expression of *Plasmodium falciparum* malaria in rural Amazonians. *Am J Trop Med Hyg*, 76, 1084-91.
- SCOPEL, K. K., FONTES, C. J., FERREIRA, M. U. & BRAGA, E. M. 2005. *Plasmodium falciparum*: IgG subclass antibody response to merozoite surface protein-1 among Amazonian gold miners, in relation to infection status and disease expression. *Exp Parasitol*, 109, 124-34.
- SEDER, R. A., CHANG, L. J., ENAMA, M. E., ZEPHIR, K. L., SARWAR, U. N., GORDON, I. J., HOLMAN, L. A., JAMES, E. R., BILLINGSLEY, P. F., GUNASEKERA, A., RICHMAN, A., CHAKRAVARTY, S., MANOJ, A., VELMURUGAN, S., LI, M., RUBEN, A. J., LI, T., EAPPEN, A. G., STAFFORD, R. E., PLUMMER, S. H., HENDEL, C. S., NOVIK, L., COSTNER, P. J., MENDOZA, F. H., SAUNDERS, J. G., NASON, M. C., RICHARDSON, J. H., MURPHY, J., DAVIDSON, S. A., RICHIE, T. L., SEDEGAH, M., SUTAMIHARDJA, A., FAHLE, G. A., LYKE, K. E., LAURENS, M. B., ROEDERER, M., TEWARI, K., EPSTEIN, J. E., SIM, B. K., LEDGERWOOD, J. E., GRAHAM, B. S. & HOFFMAN, S. L. 2013. Protection against malaria by intravenous immunization with a nonreplicating sporozoite vaccine. *Science*, 341, 1359-65.
- SEMPERTEGUI, F., ESTRELLA, B., MOSCOSO, J., PIEDRAHITA, L., HERNANDEZ, D., GAYBOR, J., NARANJO, P., MANCERO, O., ARIAS, S., BERNAL, R. & ET AL. 1994. Safety, immunogenicity and protective effect of the SPf66 malaria synthetic vaccine against *Plasmodium falciparum* infection in a randomized double-blind placebo-controlled field trial in an endemic area of Ecuador. *Vaccine*, 12, 337-42.
- SERENE, L. 2015. *Characterization of immunoglobulin heavy and light chain variable regions from single B-cells* MSc MSc Project Report, London School of Hygiene & Tropical Medicine.
- SHEEHY, S. H., DUNCAN, C. J., ELIAS, S. C., CHOUDHARY, P., BISWAS, S., HALSTEAD, F. D., COLLINS, K. A., EDWARDS, N. J., DOUGLAS, A. D., ANAGNOSTOU, N. A., EWER, K. J., HAVELOCK, T., MAHUNGU, T., BLISS, C. M., MIURA, K., POULTON, I. D., LILLIE, P. J., ANTROBUS, R. D., BERRIE, E., MOYLE, S., GANTLETT, K., COLLOCA, S., CORTESE, R., LONG, C. A., SINDEN, R. E., GILBERT, S. C., LAWRIE, A. M., DOHERTY, T., FAUST, S. N., NICOSIA, A., HILL, A. V. & DRAPER, S. J. 2012. ChAd63-MVA-vectored blood-stage malaria vaccines targeting MSP1 and AMA1: assessment of efficacy against mosquito bite challenge in humans. *Mol Ther*, 20, 2355-68.
- SHEEHY, S. H., DUNCAN, C. J., ELIAS, S. C., COLLINS, K. A., EWER, K. J., SPENCER, A. J., WILLIAMS, A. R., HALSTEAD, F. D., MORETZ, S. E., MIURA, K., EPP, C., DICKS, M. D., POULTON, I. D., LAWRIE, A. M., BERRIE, E., MOYLE, S., LONG, C. A., COLLOCA, S., CORTESE, R., GILBERT, S. C., NICOSIA, A., HILL, A. V. & DRAPER, S. J. 2011. Phase Ia clinical evaluation of the *Plasmodium falciparum* blood-stage antigen MSP1 in ChAd63 and MVA vaccine vectors. *Mol Ther*, 19, 2269-76.
- SHIN, E. K., MATSUDA, F., NAGAOKA, H., FUKITA, Y., IMAI, T., YOKOYAMA, K., SOEDA, E. & HONJO, T. 1991. Physical map of the 3' region of the human immunoglobulin heavy chain locus: clustering of autoantibody-related variable segments in one haplotype. *EMBO J*, 10, 3641-5.
- SIEVERS, F., WILM, A., DINEEN, D., GIBSON, T. J., KARPLUS, K., LI, W., LOPEZ, R., MCWILLIAM, H., REMMERT, M., SODING, J., THOMPSON, J. D. & HIGGINS, D. G. 2011. Fast, scalable generation of high-quality protein multiple sequence alignments using Clustal Omega. *Mol Syst Biol*, 7, 539.

- SILVA, N. S., SILVEIRA, L. A., MACHADO, R. L., POVOA, M. M. & FERREIRA, M. U. 2000. Temporal and spatial distribution of the variants of merozoite surface protein-1 (MSP-1) in *Plasmodium falciparum* populations in Brazil. *Ann Trop Med Parasitol*, 94, 675-88.
- SIM, B., CHITNIS, C., WASNIEWSKA, K., HADLEY, T. & MILLER, L. 1994. Receptor and ligand domains for invasion of erythrocytes by *Plasmodium falciparum*. *Science*, 264, 1941-1944.
- SINGH, S., SOE, S., WEISMAN, S., BARNWELL, J. W., PERIGNON, J. L. & DRUILHE, P. 2009. A conserved multi-gene family induces cross-reactive antibodies effective in defense against *Plasmodium falciparum*. *PLoS One*, 4, e5410.
- SIRIMA, S. B., COUSENS, S. & DRUILHE, P. 2011. Protection against malaria by MSP3 candidate vaccine. *N Engl J Med*, 365, 1062-4.
- SIRIMA, S. B., MORDMULLER, B., MILLIGAN, P., NGOA, U. A., KIRONDE, F., ATUGUBA, F., TIONO, A. B., ISSIFOU, S., KADDUMUKASA, M., BANGRE, O., FLACH, C., CHRISTIANSEN, M., BANG, P., CHILENGI, R., JEPSEN, S., KREMSNER, P. G. & THEISEN, M. 2016. A phase 2b randomized, controlled trial of the efficacy of the GM22 malaria vaccine in African children. *Vaccine*, 34, 4536-42.
- SIRIMA, S. B., NEBIE, I., OUEDRAOGO, A., TIONO, A. B., KONATE, A. T., GANSANE, A., DERME, A. I., DIARRA, A., OUEDRAOGO, A., SOULAMA, I., CUZZIN-OUATTARA, N., COUSENS, S. & LEROY, O. 2007. Safety and immunogenicity of the *Plasmodium falciparum* merozoite surface protein-3 long synthetic peptide (MSP3-LSP) malaria vaccine in healthy, semi-immune adult males in Burkina Faso, West Africa. *Vaccine*, 25, 2723-32.
- SISKIND, G. W. & BENACERRAF, B. 1969. Cell selection by antigen in the immune response. *Adv Immunol*, 10, 1-50.
- SMYTHE, J. A., PETERSON, M. G., COPPEL, R. L., SAUL, A. J., KEMP, D. J. & ANDERS, R. F. 1990. Structural diversity in the 45-kilodalton merozoite surface antigen of *Plasmodium falciparum*. *Mol Biochem Parasitol*, 39, 227-34.
- SOE, S., SINGH, S., CAMUS, D., HORII, T. & DRUILHE, P. 2002. *Plasmodium falciparum* serine repeat protein, a new target of monocyte-dependent antibody-mediated parasite killing. *Infect Immun*, 70, 7182-4.
- SOE, S., THEISEN, M., ROUSSILHON, C., AYE, K.-S.-. & DRUILHE, P. 2004. Association between Protection against Clinical Malaria and Antibodies to Merozoite Surface Antigens in an Area of Hyperendemicity in Myanmar: Complementarity between Responses to Merozoite Surface Protein 3 and the 220-Kilodalton Glutamate-Rich Protein. *Infection and Immunity*, 72, 247-252.
- SOWA, K. M., CAVANAGH, D. R., CREASEY, A. M., RAATS, J., MCBRIDE, J., SAUERWEIN, R., ROEFFEN, W. F. & ARNOT, D. E. 2001. Isolation of a monoclonal antibody from a malaria patient-derived phage display library recognising the Block 2 region of *Plasmodium falciparum* merozoite surface protein-1. *Mol Biochem Parasitol*, 112, 143-7.
- SRINIVASAN, P., BEATTY, W. L., DIOUF, A., HERRERA, R., AMBROGGIO, X., MOCH, J. K., TYLER, J. S., NARUM, D. L., PIERCE, S. K., BOOTHROYD, J. C., HAYNES, J. D. & MILLER, L. H. 2011. Binding of *Plasmodium* merozoite proteins RON2 and AMA1 triggers commitment to invasion. *Proc Natl Acad Sci U S A*, 108, 13275-80.
- STADEN, R. 1979. A strategy of DNA sequencing employing computer programs. *Nucleic Acids Res*, 6, 2601-10.
- STAFFORD, W. H., BLACKMAN, M. J., HARRIS, A., SHAI, S., GRAINGER, M. & HOLDER, A. A. 1994. N-terminal amino acid sequence of the *Plasmodium falciparum* merozoite surface protein-1 polypeptides. *Mol Biochem Parasitol*, 66, 157-60.
- STANISIC, D. I., RICHARDS, J. S., MCCALLUM, F. J., MICHON, P., KING, C. L., SCHOEPLIN, S., GILSON, P. R., MURPHY, V. J., ANDERS, R. F., MUELLER, I. & BEESON, J. G. 2009. Immunoglobulin G subclass-specific responses against *Plasmodium falciparum* merozoite antigens are associated with control of parasitemia and protection from symptomatic illness. *Infect Immun*, 77, 1165-74.

- STEWART, L., GOSLING, R., GRIFFIN, J., GESASE, S., CAMPO, J., HASHIM, R., MASIKA, P., MOSHA, J., BOUSEMA, T., SHEKALAGHE, S., COOK, J., CORRAN, P., GHANI, A., RILEY, E. M. & DRAKELEY, C. 2009. Rapid assessment of malaria transmission using age-specific sero-conversion rates. *PLoS One*, 4, e6083.
- STONE, W. J., ELDERING, M., VAN GEMERT, G. J., LANKE, K. H., GRIGNARD, L., VAN DE VEGTE-BOLMER, M. G., SIEBELINK-STOTER, R., GRAUMANS, W., ROEFFEN, W. F., DRAKELEY, C. J., SAUERWEIN, R. W. & BOUSEMA, T. 2013. The relevance and applicability of oocyst prevalence as a read-out for mosquito feeding assays. *Sci Rep*, 3, 3418.
- STOUTE, J. A., GOMBE, J., WITHERS, M. R., SIANGLA, J., MCKINNEY, D., ONYANGO, M., CUMMINGS, J. F., MILMAN, J., TUCKER, K., SOISSON, L., STEWART, V. A., LYON, J. A., ANGOV, E., LEACH, A., COHEN, J., KESTER, K. E., OCKENHOUSE, C. F., HOLLAND, C. A., DIGGS, C. L., WITTES, J. & HEPPNER, D. G., JR. 2007. Phase 1 randomized double-blind safety and immunogenicity trial of Plasmodium falciparum malaria merozoite surface protein FMP1 vaccine, adjuvanted with AS02A, in adults in western Kenya. *Vaccine*, 25, 176-84.
- STUBBS, J., OLUGBILE, S., SAIDOU, B., SIMPORE, J., CORRADIN, G. & LANZAVECCHIA, A. 2011. Strain-Transcending Fc-Dependent Killing of Plasmodium falciparum by Merozoite Surface Protein 2 Allele-Specific Human Antibodies. *Infection and Immunity*, 79, 1143-1152.
- STUDIER, F. W. 2005. Protein production by auto-induction in high density shaking cultures. *Protein Expr Purif*, 41, 207-34.
- STURCHLER, D., BERGER, R., RUDIN, C., JUST, M., SAUL, A., RZEPCHYK, C., BROWN, G., ANDERS, R., COPPEL, R., WOODROW, G. & ET AL. 1995. Safety, immunogenicity, and pilot efficacy of Plasmodium falciparum sporozoite and asexual blood-stage combination vaccine in Swiss adults. *Am J Trop Med Hyg*, 53, 423-31.
- SU, S., SANADI, A. R., IFON, E. & DAVIDSON, E. A. 1993. A monoclonal antibody capable of blocking the binding of Pf200 (MSA-1) to human erythrocytes and inhibiting the invasion of Plasmodium falciparum merozoites into human erythrocytes. *J Immunol*, 151, 2309-17.
- SWINDELLS, M. B., PORTER, C. T., COUCH, M., HURST, J., ABHINANDAN, K. R., NIELSEN, J. H., MACINDOE, G., HETHERINGTON, J. & MARTIN, A. C. 2017. abYsis: Integrated Antibody Sequence and Structure-Management, Analysis, and Prediction. *J Mol Biol*, 429, 356-364.
- TAKALA, S., BRANCH, O., ESCALANTE, A. A., KARIUKI, S., WOOTTON, J. & LAL, A. A. 2002. Evidence for intragenic recombination in Plasmodium falciparum: identification of a novel allele family in block 2 of merozoite surface protein-1: Asembo Bay Area Cohort Project XIV. *Mol Biochem Parasitol*, 125, 163-71.
- TAKALA, S. L., ESCALANTE, A. A., BRANCH, O. H., KARIUKI, S., BISWAS, S., CHAIYAROJ, S. C. & LAL, A. A. 2006. Genetic diversity in the Block 2 region of the merozoite surface protein 1 (MSP-1) of Plasmodium falciparum: additional complexity and selection and convergence in fragment size polymorphism. *Infect Genet Evol*, 6, 417-24.
- TAKALA, S. L. & PLOWE, C. V. 2009. Genetic diversity and malaria vaccine design, testing and efficacy: preventing and overcoming 'vaccine resistant malaria'. *Parasite Immunol*, 31, 560-73.
- TAN, J., PIEPER, K., PICCOLI, L., ABDI, A., FOGLIERINI, M., GEIGER, R., TULLY, C. M., JARROSSAY, D., NDUNGU, F. M., WAMBUA, J., BEJON, P., FREGNI, C. S., FERNANDEZ-RODRIGUEZ, B., BARBIERI, S., BIANCHI, S., MARSH, K., THATHY, V., CORTI, D., SALLUSTO, F., BULL, P. & LANZAVECCHIA, A. 2016. A LAIR1 insertion generates broadly reactive antibodies against malaria variant antigens. *Nature*, 529, 105-9.
- TANABE, K., MACKAY, M., GOMAN, M. & SCAIFE, J. G. 1987. Allelic dimorphism in a surface antigen gene of the malaria parasite Plasmodium falciparum. *Journal of Molecular Biology*, 195, 273-287.
- TANABE, K., MITA, T., JOMBART, T., ERIKSSON, A., HORIBE, S., PALACPAC, N., RANFORD-CARTWRIGHT, L., SAWAI, H., SAKIHAMA, N., OHMAE, H., NAKAMURA, M., FERREIRA, M. U., ESCALANTE, A. A., PRUGNOLLE, F., BJORKMAN, A., FARNERT, A., KANEKO, A., HORII, T.,

- MANICA, A., KISHINO, H. & BALLOUX, F. 2010. Plasmodium falciparum accompanied the human expansion out of Africa. *Curr Biol*, 20, 1283-9.
- TANABE, K., MITA, T., PALACPAC, N. M., ARISUE, N., TOUGAN, T., KAWAI, S., JOMBART, T., KOBAYASHI, F. & HORII, T. 2013. Within-population genetic diversity of Plasmodium falciparum vaccine candidate antigens reveals geographic distance from a Central sub-Saharan African origin. *Vaccine*, 31, 1334-9.
- TANABE, K., SAKIHAMA, N., ROTH, I., BJORKMAN, A. & FARNERT, A. 2007a. High frequency of recombination-driven allelic diversity and temporal variation of Plasmodium falciparum msp1 in Tanzania. *Am J Trop Med Hyg*, 76, 1037-45.
- TANABE, K., SAKIHAMA, N., WALLIKER, D., BABIKER, H., ABDEL-MUHSIN, A. M., BAKOTE'E, B., OHMAE, H., ARISUE, N., HORII, T., ROTH, I., FARNERT, A., BJORKMAN, A. & RANFORD-CARTWRIGHT, L. 2007b. Allelic dimorphism-associated restriction of recombination in Plasmodium falciparum msp1. *Gene*, 397, 153-60.
- TANABE, K., ZOLLNER, G. E., SATTABONGKOT, J., KHUNTIRAT, B., HONMA, H., MITA, T., TSUBOI, T. AND COLEMAN, R. 2013. Genetic diversity of Plasmodium falciparum in an isolated village in western Thailand. *Unpublished*.
- TARASOV, A., VILELLA, A. J., CUPPEN, E., NIJMAN, I. J. & PRINS, P. 2015. Sambamba: fast processing of NGS alignment formats. *Bioinformatics*, 31, 2032-4.
- TARLINTON, D. 2006. B-cell memory: are subsets necessary? *Nat Rev Immunol*, 6, 785-90.
- TAYLOR, R. R., ALLEN, S. J., GREENWOOD, B. M. & RILEY, E. M. 1998. IgG3 antibodies to Plasmodium falciparum merozoite surface protein 2 (MSP2): increasing prevalence with age and association with clinical immunity to malaria. *Am J Trop Med Hyg*, 58, 406-13.
- TAYLOR, R. R., SMITH, D. B., ROBINSON, V. J., MCBRIDE, J. S. & RILEY, E. M. 1995. Human antibody response to Plasmodium falciparum merozoite surface protein 2 is serogroup specific and predominantly of the immunoglobulin G3 subclass. *Infection and Immunity*, 63, 4382-8.
- TEAM, R. D. C. (ed.) 2008. *R: A language and environment for statistical computing*, Vienna, Austria: R Foundation for Statistical Computing.
- TETTEH, K. K., CAVANAGH, D. R., CORRAN, P., MUSONDA, R., MCBRIDE, J. S. & CONWAY, D. J. 2005a. Extensive antigenic polymorphism within the repeat sequence of the Plasmodium falciparum merozoite surface protein 1 block 2 is incorporated in a minimal polyvalent immunogen. *Infect Immun*, 73, 5928-35.
- TETTEH, K. K. & CONWAY, D. J. 2011. A polyvalent hybrid protein elicits antibodies against the diverse allelic types of block 2 in Plasmodium falciparum merozoite surface protein 1. *Vaccine*, 29, 7811-7.
- TETTEH, K. K., STEWART, L. B., OCHOLA, L. I., AMAMBUA-NGWA, A., THOMAS, A. W., MARSH, K., WEEDALL, G. D. & CONWAY, D. J. 2009. Prospective Identification of Malaria Parasite Genes under Balancing Selection. *PLoS ONE*, 4, e5568.
- TETTEH, K. K. A., CAVANAGH, D. R., CORRAN, P., MUSONDA, R., MCBRIDE, J. S. & CONWAY, D. J. 2005b. Extensive Antigenic Polymorphism within the Repeat Sequence of the Plasmodium falciparum Merozoite Surface Protein 1 Block 2 Is Incorporated in a Minimal Polyvalent Immunogen. *Infection and Immunity*, 73, 5928-5935.
- THAM, W. H., HEALER, J. & COWMAN, A. F. 2012. Erythrocyte and reticulocyte binding-like proteins of Plasmodium falciparum. *Trends Parasitol*, 28, 23-30.
- THAM, W. H., SCHMIDT, C. Q., HAUHART, R. E., GUARIENTO, M., TETTEH-QUARCOO, P. B., LOPATICKI, S., ATKINSON, J. P., BARLOW, P. N. & COWMAN, A. F. 2011. Plasmodium falciparum uses a key functional site in complement receptor type-1 for invasion of human erythrocytes. *Blood*, 118, 1923-33.
- THEISEN, M., SOE, S., BRUNSTEDT, K., FOLLMANN, F., BREDMOSE, L., ISRAESEN, H., MADSEN, S. M. & DRUILHE, P. 2004. A Plasmodium falciparum GLURP-MSP3 chimeric protein; expression in Lactococcus lactis, immunogenicity and induction of biologically active antibodies. *Vaccine*, 22, 1188-98.

- THEISEN, M., SOE, S., OEUVRAY, C., THOMAS, A. W., VUUST, J., DANIELSEN, S., JEPSEN, S. & DRUILHE, P. 1998. The glutamate-rich protein (GLURP) of *Plasmodium falciparum* is a target for antibody-dependent monocyte-mediated inhibition of parasite growth in vitro. *Infect Immun*, 66, 11-7.
- THERA, M. A., DOUMBO, O. K., COULIBALY, D., DIALLO, D. A., KONE, A. K., GUINDO, A. B., TRAORE, K., DICKO, A., SAGARA, I., SISSOKO, M. S., BABY, M., SISSOKO, M., DIARRA, I., NIANGALY, A., DOLO, A., DAOU, M., DIAWARA, S. I., HEPPNER, D. G., STEWART, V. A., ANGOV, E., BERGMANN-LEITNER, E. S., LANAR, D. E., DUTTA, S., SOISSON, L., DIGGS, C. L., LEACH, A., OWUSU, A., DUBOIS, M. C., COHEN, J., NIXON, J. N., GREGSON, A., TAKALA, S. L., LYKE, K. E. & PLOWE, C. V. 2008. Safety and immunogenicity of an AMA-1 malaria vaccine in Malian adults: results of a phase 1 randomized controlled trial. *PLoS One*, 3, e1465.
- THERA, M. A., DOUMBO, O. K., COULIBALY, D., DIALLO, D. A., SAGARA, I., DICKO, A., DIEMERT, D. J., HEPPNER, D. G., JR., STEWART, V. A., ANGOV, E., SOISSON, L., LEACH, A., TUCKER, K., LYKE, K. E. & PLOWE, C. V. 2006. Safety and allele-specific immunogenicity of a malaria vaccine in Malian adults: results of a phase I randomized trial. *PLoS Clin Trials*, 1, e34.
- THERA, M. A., DOUMBO, O. K., COULIBALY, D., LAURENS, M. B., OUATTARA, A., KONE, A. K., GUINDO, A. B., TRAORE, K., TRAORE, I., KOURIBA, B., DIALLO, D. A., DIARRA, I., DAOU, M., DOLO, A., TOLO, Y., SISSOKO, M. S., NIANGALY, A., SISSOKO, M., TAKALA-HARRISON, S., LYKE, K. E., WU, Y., BLACKWELDER, W. C., GODEAUX, O., VEKEMANS, J., DUBOIS, M. C., BALLOU, W. R., COHEN, J., THOMPSON, D., DUBE, T., SOISSON, L., DIGGS, C. L., HOUSE, B., LANAR, D. E., DUTTA, S., HEPPNER, D. G., JR. & PLOWE, C. V. 2011. A field trial to assess a blood-stage malaria vaccine. *N Engl J Med*, 365, 1004-13.
- THOMPSON, F. M., PORTER, D. W., OKITSU, S. L., WESTERFELD, N., VOGEL, D., TODRYK, S., POULTON, I., CORREA, S., HUTCHINGS, C., BERTHOUD, T., DUNACHIE, S., ANDREWS, L., WILLIAMS, J. L., SINDEN, R., GILBERT, S. C., PLUSCHKE, G., ZURBRIGGEN, R. & HILL, A. V. 2008. Evidence of blood stage efficacy with a virosomal malaria vaccine in a phase IIa clinical trial. *PLoS One*, 3, e1493.
- TILLER, T., MEFFRE, E., YURASOV, S., TSUIJI, M., NUSSENZWEIG, M. C. & WARDEMAN, H. 2008. Efficient generation of monoclonal antibodies from single human B cells by single cell RT-PCR and expression vector cloning. *Journal of Immunological Methods*, 329, 112-124.
- TINE, J. A., LANAR, D. E., SMITH, D. M., WELLDE, B. T., SCHULTHEISS, P., WARE, L. A., KAUFFMAN, E. B., WIRTZ, R. A., DE TAISSNE, C., HUI, G. S., CHANG, S. P., CHURCH, P., HOLLINGDALE, M. R., KASLOW, D. C., HOFFMAN, S., GUITO, K. P., BALLOU, W. R., SADOFF, J. C. & PAOLETTI, E. 1996. NYVAC-Pf7: a poxvirus-vectored, multiantigen, multistage vaccine candidate for *Plasmodium falciparum* malaria. *Infect Immun*, 64, 3833-44.
- TOWNSEND, S. E., GOODNOW, C. C. & CORNALL, R. J. 2001. Single epitope multiple staining to detect ultralow frequency B cells. *J Immunol Methods*, 249, 137-46.
- TRAGGIAI, E. 2012. immortalization of Human B Cells: Analysis of B Cell Repertoire and Production of Human Monoclonal Antibodies. In: PROETZEL, G. & EBERSBACH, H. (eds.) *Antibody Methods and Protocols*. Humana Press.
- TRAN, T. M., ONGOIBA, A., COURSEN, J., CROSNIER, C., DIOUF, A., HUANG, C. Y., LI, S., DOUMBO, S., DOUMTABE, D., KONE, Y., BATHILY, A., DIA, S., NIANGALY, M., DARA, C., SANGALA, J., MILLER, L. H., DOUMBO, O. K., KAYENTAO, K., LONG, C. A., MIURA, K., WRIGHT, G. J., TRAORE, B. & CROMPTON, P. D. 2014. Naturally acquired antibodies specific for *Plasmodium falciparum* reticulocyte-binding protein homologue 5 inhibit parasite growth and predict protection from malaria. *J Infect Dis*, 209, 789-98.
- TRKOLA, A., PURTSCHER, M., MUSTER, T., BALLAUN, C., BUCHACHER, A., SULLIVAN, N., SRINIVASAN, K., SODROSKI, J., MOORE, J. P. & KATINGER, H. 1996. Human monoclonal antibody 2G12 defines a distinctive neutralization epitope on the gp120 glycoprotein of human immunodeficiency virus type 1. *J Virol*, 70, 1100-8.

- TRUCCO, C., FERNANDEZ-REYES, D., HOWELL, S., STAFFORD, W. H., SCOTT-FINNIGAN, T. J., GRAINGER, M., OGUN, S. A., TAYLOR, W. R. & HOLDER, A. A. 2001. The merozoite surface protein 6 gene codes for a 36 kDa protein associated with the Plasmodium falciparum merozoite surface protein-1 complex. *Mol Biochem Parasitol*, 112, 91-101.
- TURNER, L., WANG, C. W., LAVSTSEN, T., MWAKALINGA, S. B., SAUERWEIN, R. W., HERMSEN, C. C. & THEANDER, T. G. 2011. Antibodies against PfEMP1, RIFIN, MSP3 and GLURP are acquired during controlled Plasmodium falciparum malaria infections in naive volunteers. *PLoS One*, 6, e29025.
- UDEY, J. A. & BLOMBERG, B. 1987. Human lambda light chain locus: organization and DNA sequences of three genomic J regions. *Immunogenetics*, 25, 63-70.
- UDOMSANGPETCH, R., LUNDGREN, K., BERZINS, K., WAHLIN, B., PERLMANN, H., TROYE-BLOMBERG, M., CARLSSON, J., WAHLGREN, M., PERLMANN, P. & BJORKMAN, A. 1986. Human monoclonal antibodies to Pf 155, a major antigen of malaria parasite Plasmodium falciparum. *Science*, 231, 57-9.
- UDUMAN, M., YAARI, G., HERSHBERG, U., STERN, J. A., SHLOMCHIK, M. J. & KLEINSTEIN, S. H. 2011. Detecting selection in immunoglobulin sequences. *Nucleic Acids Res*, 39, W499-504.
- URDANETA, M., PRATA, A., STRUCHINER, C. J., TOSTA, C. E., TAUIL, P. & BOULOS, M. 1998. Evaluation of SPf66 malaria vaccine efficacy in Brazil. *Am J Trop Med Hyg*, 58, 378-85.
- VALERO, M. V., AMADOR, L. R., GALINDO, C., FIGUEROA, J., BELLO, M. S., MURILLO, L. A., MORA, A. L., PATARROYO, G., ROCHA, C. L., ROJAS, M. & ET AL. 1993. Vaccination with SPf66, a chemically synthesised vaccine, against Plasmodium falciparum malaria in Colombia. *Lancet*, 341, 705-10.
- VALERO, M. V., AMADOR, R., APONTE, J. J., NARVAEZ, A., GALINDO, C., SILVA, Y., ROSAS, J., GUZMAN, F. & PATARROYO, M. E. 1996. Evaluation of SPf66 malaria vaccine during a 22-month follow-up field trial in the Pacific coast of Colombia. *Vaccine*, 14, 1466-70.
- VEMBAR, S. S., SEETIN, M., LAMBERT, C., NATTESTAD, M., SCHATZ, M. C., BAYBAYAN, P., SCHERF, A. & SMITH, M. L. 2016. Complete telomere-to-telomere de novo assembly of the Plasmodium falciparum genome through long-read (>11 kb), single molecule, real-time sequencing. *DNA Res*, 23, 339-51.
- VIGAN-WOMAS, I., GUILLOTTE, M., JUILLERAT, A., HESSEL, A., RAYNAL, B., ENGLAND, P., COHEN, J. H., BERTRAND, O., PEYRARD, T., BENTLEY, G. A., LEWIT-BENTLEY, A. & MERCEREAU-PUIJALON, O. 2012. Structural basis for the ABO blood-group dependence of Plasmodium falciparum rosetting. *PLoS Pathog*, 8, e1002781.
- VILLASIS, E., LOPEZ-PEREZ, M., TORRES, K., GAMBOA, D., NEYRA, V., BENDEZU, J., TRICOCHÉ, N., LOBO, C., VINETZ, J. M. & LUSTIGMAN, S. 2012. Anti-Plasmodium falciparum invasion ligand antibodies in a low malaria transmission region, Loreto, Peru. *Malar J*, 11, 361.
- VIRIYAKOSOL, S., SIRIPOON, N., PETCHARAPIRAT, C., PETCHARAPIRAT, P., JARRA, W., THAITHONG, S., BROWN, K. N. & SNOUNOU, G. 1995. Genotyping of Plasmodium falciparum isolates by the polymerase chain reaction and potential uses in epidemiological studies. *Bull World Health Organ*, 73, 85-95.
- VOLKMAN, S. K., HARTL, D. L., WIRTH, D. F., NIELSEN, K. M., CHOI, M., BATALOV, S., ZHOU, Y., PLOUFFE, D., LE ROCH, K. G., ABAGYAN, R. & WINZELER, E. A. 2002. Excess polymorphisms in genes for membrane proteins in Plasmodium falciparum. *Science*, 298, 216-8.
- WAHLIN, B., WAHLGREN, M., PERLMANN, H., BERZINS, K., BJORKMAN, A., PATARROYO, M. E. & PERLMANN, P. 1984. Human antibodies to a Mr 155,000 Plasmodium falciparum antigen efficiently inhibit merozoite invasion. *Proc Natl Acad Sci U S A*, 81, 7912-6.
- WANAGURU, M., LIU, W., HAHN, B. H., RAYNER, J. C. & WRIGHT, G. J. 2013. RH5-Basigin interaction plays a major role in the host tropism of Plasmodium falciparum. *Proc Natl Acad Sci U S A*, 110, 20735-40.
- WARDEMANN, H. & KOFER, J. 2013. Expression cloning of human B cell immunoglobulins. *Methods Mol Biol*, 971, 93-111.

- WATSON, C. T., STEINBERG, K. M., HUDDLESTON, J., WARREN, R. L., MALIG, M., SCHEIN, J., WILLSEY, A. J., JOY, J. B., SCOTT, J. K., GRAVES, T. A., WILSON, R. K., HOLT, R. A., EICHLER, E. E. & BREDEN, F. 2013. Complete haplotype sequence of the human immunoglobulin heavy-chain variable, diversity, and joining genes and characterization of allelic and copy-number variation. *Am J Hum Genet*, 92, 530-46.
- WEAVER, R., REILING, L., FENG, G., DREW, D. R., MUELLER, I., SIBA, P. M., TSUBOI, T., RICHARDS, J. S., FOWKES, F. J. & BEESON, J. G. 2016. The association between naturally acquired IgG subclass specific antibodies to the PfrH5 invasion complex and protection from *Plasmodium falciparum* malaria. *Sci Rep*, 6, 33094.
- WEEDALL, G. D. & CONWAY, D. J. 2010. Detecting signatures of balancing selection to identify targets of anti-parasite immunity. *Trends in Parasitology*, 26, 363-369.
- WHO 2015. World Malaria Report 2015. Geneva, Switzerland: World Health Organization.
- WHO 2016. Malaria Vaccine: WHO position paper - January 2016. *Weekly Epidemiological Record*, 91, 33-52.
- WICKHAM, H. 2009. ggplot2: elegant graphics for data analysis. Springer New York.
- WILLIAMS, S. C., FRIPPIAT, J. P., TOMLINSON, I. M., IGNATOVICH, O., LEFRANC, M. P. & WINTER, G. 1996. Sequence and evolution of the human germline V lambda repertoire. *J Mol Biol*, 264, 220-32.
- WILSON, R. J. & PHILLIPS, R. S. 1976. Method to test inhibitory antibodies in human sera to wild populations of *Plasmodium falciparum*. *Nature*, 263, 132-4.
- WITHERS, M. R., MCKINNEY, D., OGUTU, B. R., WAITUMBI, J. N., MILMAN, J. B., APOLLO, O. J., ALLEN, O. G., TUCKER, K., SOISSON, L. A., DIGGS, C., LEACH, A., WITTES, J., DUBOVSKY, F., STEWART, V. A., REMICH, S. A., COHEN, J., BALLOU, W. R., HOLLAND, C. A., LYON, J. A., ANGOV, E., STOUTE, J. A., MARTIN, S. K. & HEPPNER, D. G., JR. 2006. Safety and reactogenicity of an MSP-1 malaria vaccine candidate: a randomized phase Ib dose-escalation trial in Kenyan children. *PLoS Clin Trials*, 1, e32.
- WOFSY, L. & BURR, B. 1969. The use of affinity chromatography for the specific purification of antibodies and antigens. *J Immunol*, 103, 380-2.
- XU, J. L. & DAVIS, M. M. 2000. Diversity in the CDR3 region of V(H) is sufficient for most antibody specificities. *Immunity*, 13, 37-45.
- YAARI, G., UDUMAN, M. & KLEINSTEIN, S. H. 2012. Quantifying selection in high-throughput Immunoglobulin sequencing data sets. *Nucleic Acids Res*, 40, e134.
- YAGI, M., PALACPAC, N. M., ITO, K., OISHI, Y., ITAGAKI, S., BALIKAGALA, B., NTEGE, E. H., YEKA, A., KANOI, B. N., KATURO, O., SHIRAI, H., FUKUSHIMA, W., HIROTA, Y., EGWANG, T. G. & HORII, T. 2016. Antibody titres and boosting after natural malaria infection in BK-SE36 vaccine responders during a follow-up study in Uganda. *Sci Rep*, 6, 34363.
- YE, J., MA, N., MADDEN, T. L. & OSTELL, J. M. 2013. IgBLAST: an immunoglobulin variable domain sequence analysis tool. *Nucleic Acids Res*, 41, W34-40.
- YUTHAVONG, Y., BUNYARATVEJ, A. & KAMCHONWONGPAISAN, S. 1990. Increased susceptibility of malaria-infected variant erythrocytes to the mononuclear phagocyte system. *Blood Cells*, 16, 591-7.
- ZERBINO, D. R. 2010. Using the Velvet de novo assembler for short-read sequencing technologies. *Curr Protoc Bioinformatics*, Chapter 11, Unit 11.5.
- ZERBINO, D. R. & BIRNEY, E. 2008. Velvet: algorithms for de novo short read assembly using de Bruijn graphs. *Genome Res*, 18, 821-9.
- ZHANG, W., CHEN, J., YANG, Y., TANG, Y., SHANG, J. & SHEN, B. 2011. A practical comparison of de novo genome assembly software tools for next-generation sequencing technologies. *PLoS One*, 6, e17915.
- ZHANG, Y., JIANG, N., LU, H., HOU, N., PIAO, X., CAI, P., YIN, J., WAHLGREN, M. & CHEN, Q. 2013. Proteomic analysis of *Plasmodium falciparum* schizonts reveals heparin-binding merozoite proteins. *J Proteome Res*, 12, 2185-93.

7. Appendices

7.1 *Msp1* block 2 genotyping studies

Region	Study	K1	MAD20	RO-33	MR
Cental Africa	Boyoun-Akotet et al., 2015	71	64	38	N/A
	Mawili-Mboumba et al., 2015	110	35	94	N/A
	Yavo et al., 2016	30	23	17	N/A
	Apinjoh et al., 2015	113	77	64	59
	Conway et al., 2000	67	46	11	N/A
	Mayengue et al., 2011	60	36	29	N/A
	Aubouy et al., 2003	47	33	19	N/A
	Dolmazon et al., 2008	41	103	201	N/A
	Total	539	417	473	59
East Africa	Branch et al., 2001	640	544	564	N/A
	Conway et al., 2000	46	22	18	N/A
	Mohammed et al., 2015	43	15	31	N/A
	Mwingira et al., 2011	248	142	105	N/A
	Peyerl-Hoffman et al., 2001	182	93	80	N/A
	Robinson et al., 2011	3	3	2	N/A
	Takala et al., 2002	168	131	147	48
	Takala et al., 2006	344	261	286	105
	Tanabe et al., 2007	388	218	222	N/A
	Jiang et al., 2000	17	6	4	N/A
	Juliano et al., 2010	48	18	11	N/A
	Total	2127	1453	1470	153
North Africa	Conway et al., 2000	24	22	20	N/A
	Mahdi Abdel Hamid et al., 2016	57	71	71	N/A
	Hamid et al., 2013	12	15	16	N/A
	Total	93	108	107	
Southern Africa	Conway et al., 2000	43	20	10	N/A
	Total	43	20	10	N/A

Region	Study	K1	MAD20	RO-33	MR
West Africa	Ayanful-Torgby et al., 2016	62	60	52	N/A
	Bamidele Abiodun et al., 2014	34	12	7	N/A
	Conway et al., 2000	102	61	35	N/A
	Liljander et al., 2009	197	129	126	N/A
	Mwingira et al., 2011	171	98	102	N/A
	Niang et al., 2016	83	104	71	N/A
	Niang et al., 2017	143	140	100	N/A
	Ogouyèmi-Hounto, Gazard et al., 2013	75	59	73	N/A
	Ogouyèmi-Hounto, Ndam et al., 2013	175	129	130	N/A
	Kolawole et al., 2016	27	20	23	N/A
	Robinson et al., 2011	6	2	2	N/A
	Tanabe et al., 2010	20	8	3	1
	Ahmedou Salem et al., 2014	102	77	74	N/A
	Schleiermacher et al., 2001	173	43	151	N/A
	Soulama et al., 2009	153	119	123	N/A
	Yavo et al., 2016	53	33	29	N/A
	Noranate et al., 2009	247	145	132	22
	Scherf et al., 1991	9	2	15	N/A
	Total	1832	1241	1248	23
Africa	Total	4634	3239	3308	235
South Asia	Ghanchi et al., 2010	62	109	33	N/A
	Joshi et al., 2007	76	56	39	N/A
	Saha et al., 2012	136	90	150	N/A
	Saha et al., 2016	68	36	64	N/A
	Raj et al., 2004	71	197	140	N/A
	Total	413	488	426	N/A
South East Asia	Conway et al., 2000	142	422	91	N/A
	Gosi et al., 2013	118	7	7	N/A
	Jongwutiwes et al., 1992	6	16	4	N/A
	Juliano et al., 2010	1	12	1	N/A
	Khaminsou et al., 2011	154	107	72	N/A
	Mohd Abd Razak et al., 2016	28	18	19	N/A
	Kang et al., 2010	46	57	N/A	N/A
	Sakihama et al., 1999	9	58	11	N/A
	Sakihama et al., 2006	104	120	82	N/A
	Sakihama et al., 2007	19	37	1	N/A
	Sulistyaningsih et al., 2013	9	19	2	N/A
	Tanabe et al., 2013	9	17	0	N/A
	Yuan et al., 2013	83	115	77	N/A
	Total	728	1005	367	N/A
West Asia	Al-abd et al., 2013	33	16	31	N/A
	Total	33	16	31	N/A
Asia	Total	1174	1509	824	N/A

Region	Study	K1	MAD20	RO-33	MR
Melanesia	Sakihama et al., 2006	73	117	213	N/A
	Tanabe et al., 2010	22	49	5	N/A
	Total	95	166	218	N/A
Amazon basin	Snewin et al., 1991	3	15	26	N/A
	Gómez et al., 2002	12	66	27	N/A
	Terrientes et al, 2005	0	46	0	N/A
	Montoya et al., 2003	0	100	0	N/A
	Osorio et al., 2007a, 2007b	41	416	0	N/A
	Guerra et al., 2006	8	118	3	N/A
	Jimenez et al., 2010	11	94	6	N/A
	Kimura et al., 1990	2	1	3	N/A
	Medeiros et al., 2013	18	12	26	N/A
	Scopel et al., 2005	20	18	11	N/A
	Tanabe et al., 2009	22	13	15	N/A
	Maestre et al., 2013	20	434	9	N/A
	Zervos et al., 2012	38	38	0	N/A
	Ferreira et al., 1998	37	15	25	N/A
	Silva et al., 2000	121	79	136	N/A
	Total	195	1386	151	N/A
All	Total	6256	6379	4637	705

PubMed was searched using the terms “*plasmodium falciparum*”, “*msp1*” and “genotyping” on the 22nd March 2017, returning 85 studies. All papers were read and all studies presenting data on genotyping of *msp1* block 2 were included resulting in a total of 78 studies. The counts of K1-like, MAD20-like, RO-33-like and, where applicable, MR recombinant alleles detected are shown. Studies are categorised by region, with totals shown for each region along with totals for Africa, Asia and all studies.

7.2 Long read sequences from GenBank in the long read dataset (LRD)

Accession numbers	sequences	Reference
AB116596-AB116601	6	Tanabe et al., 2004
AB276001-AB276018, AB300615-AB300614	20	Tanabe et al., 2007b
AB502443-AB502745, AB715435-AB715519	388	Tanabe et al., 2013
AB502746-AB502795	50	Tanabe, 2009
AB827737-AB827762	26	Tanabe, 2013
AF061119-AF061151	33	Jiang et al., 2000
AF062348-AF062349	2	Jiang et al., 1999
AF218248	1	Shan, 1999
AF462449-AF462456, AY826427-AY826431, DQ377133-DQ377137	18	Takala et al., 2006
AJ635200	1	AlFadhli and Orjih, 2004
AY714585-AY714586	2	Scopel et al., 2005
DQ026701-DQ026702	2	Joshi et al., 2005
DQ447647	1	Colborn, 2006
DQ485417-DQ485451	35	Joshi et al., 2007
DQ855130-DQ855135	5	Kwiek et al., 2007
EU032016-EU037095	262	Noranate et al., 2009
HM153166-HM153256	91	Juliano et al., 2010
HQ821869-HQ821872	4	Mobassir, 2010
JF300128	1	Bharti, 2011
JX315617, JX412318-JX412322, JX416338-JX416341	9	Medeiros et al., 2013
KP318436-KP318438	3	Sehgal, 2014
KR063228-KR063231	4	Sehgal, 2015
L10380, M77713-M77737	26	Jongwutiwes et al., 1992
M19143-M19144	2	Peterson et al., 1988
M32111-M32116	6	Kimura et al., 1990
M35727/Y00087	1	Certa et al., 1987
M37213	1	Chang et al., 1988
M55001	1	Scherf et al., 1991
X02919	1	Holder et al., 1985
X03371	1	Mackay et al., 1985
X03831	1	Weber et al., 1986
X05624	1	Tanabe et al., 1987
X15063	1	Myler, 1989
X52962-X52963	2	Ranford-Cartwright et al., 1991
X61930	1	Olafsson et al., 1992
Z35327	1	Pan et al., 1995

GenBank was searched with the search terms: “*plasmodium falciparum* [organism] *msp1*”; “*plasmodium falciparum* [organism] *msa1*”; and “*plasmodium falciparum* [organism] *gp195*” on 4th December 2015. All sequences containing *msp1* block2 were downloaded. Duplicate sequences from the same laboratory strain were removed leaving 964 sequences (additional data file “long_read_sequences.fa”). The GenBank accession numbers for the sequences are shown with the number of sequences and the reference for the study for which they were produced.

7.3 Python script for generating dummy reads

The script shown below is a modification of `to_perfect_reads`, part of the `Fastaq` package downloaded on 10th November 2015 from <https://github.com/sanger-pathogens/Fastaq> and is distributed under the GNU public license, version 3, June 2007 (<https://github.com/sanger-pathogens/Fastaq/blob/master/LICENSE>). Modifications are highlighted in red.

```
def make_dummy_reads(seq_dir, out_dir, real_reads_file_basename, mean_insert, insert_std,
coverage, readlength):
    #cycles through all fasta files in a directory and creates simulated #reads from each sequence
    with quality scores from a fastq file #containing reads generated from Illumina sequencing
    #NB real read file must have reads of same length as dummy reads
    import os
    import random
    import linecache
    from math import floor, ceil
    os.mkdir(out_dir)
    real_reads1 = real_reads_file_basename+'_1.fastq'
    real_reads2 = real_reads_file_basename+'_2.fastq'
    #calculate number of reads in Illumina read file
    num_realines = sum(1 for line in open(real_reads1))
    print(int(num_realines/4), ' reads in', real_reads1)
    #create list of line numbers with quality scores

    = range(4,num_realines, 4)
    filecounter = 0
    #cycle through fasta files in directory
    for infile in os.listdir(seq_dir):
        ref=''
        rq1 = []
        rq2 = []
    #check if file is fasta
        if infile.endswith('.fa'):
            filecounter +=1
            print('Working on file number ', filecounter, ' of ', len(os.listdir(seq_dir)))
    #read in file
            with open(seq_dir+'/'+infile, 'r') as i:
                lines = i.readlines()
                for line in lines:
    #read in sequence
                    if line.startswith('>'):
                        ref_id = line.split('>')[1].strip('\n')
                        ref_id = ref_id.strip('\n')
    #get sequence ID
                    else:
                        ref=ref+line.strip('\n')
                        ref = ref.strip('\n')
    #determine fasta file names
            outfile1 =
out_dir+'/'+ref_id+str(mean_insert)+str(insert_std)+str(coverage)+str(readlength)+'.fastq1'
            outfile2 =
out_dir+'/'+ref_id+str(mean_insert)+str(insert_std)+str(coverage)+str(readlength)+'.fastq2'
    #calculate out how many dummy reads to make
            read_pairs = int(0.5 * coverage * len(ref) / readlength)
    #generate list of quality score lines (randomly picked from Illumina #reads file)
            for x in range(read_pairs):
                rndm_qual_line = random.choice(qualines)
                rq1.append(str(linecache.getline(real_reads1, rndm_qual_line)))
                rq2.append(str(linecache.getline(real_reads2, rndm_qual_line)))
    # create dictionary for recording coordinates of the read to avoid creating same read name
    #twice
            used_fragments = {}
            x = 0
            pair_counter = 1
            while x < read_pairs:
    #randomly select insert size (isize) normal distribution based on #standard deviation
    (insert_std) around mean insert size #(mean_insert)
                isize = int(random.normalvariate(mean_insert, insert_std))
    #if insert size is longer than the input sequence or shorter than #the readlength then re-
    calculate
```

```

        while isize > len(ref) or isize < readlength:
            isize = int(random.normalvariate(mean_insert, insert_std))
#calculate middle of insert (if insert size is odd, randomly select #between two middle
integers
        middle_pos = random.randint(ceil(0.5 * isize), floor(len(ref) - 0.5 * isize))
#determine read start points
        read_start1 = int(middle_pos - ceil(0.5 * isize))
        read_start2 = read_start1 + isize - readlength
#set readname
        readname = ':'.join([ref_id, str(pair_counter), str(read_start1+1),
str(read_start2+1)])
        fragment = (middle_pos, isize)
#check if fragment has been used before
        if fragment in used_fragments:
            used_fragments[fragment] += 1
#if fragment used before append "dup.x" to readname (where x = #number of times fragment is
used)
            readname += '.dup.' + str(used_fragments[fragment])
        else:
            used_fragments[fragment] = 1
#generate dummy read pair using sequence fragment and quality score #from Illumina reads file
        read1 = '@'+readname + '/1\n'+ref[read_start1:read_start1 +
readlength]+'n\n'+rq1[x]
        read2 = '@'+readname + '/2\n'+ref[read_start2:read_start2 +
readlength]+'n\n'+rq2[x]
#write out read pair
        with open(outfile1, 'a') as o:
            o.write(read1)
        with open(outfile2, 'a') as o:
            o.write(read2)
        pair_counter += 1
        x += 1

```

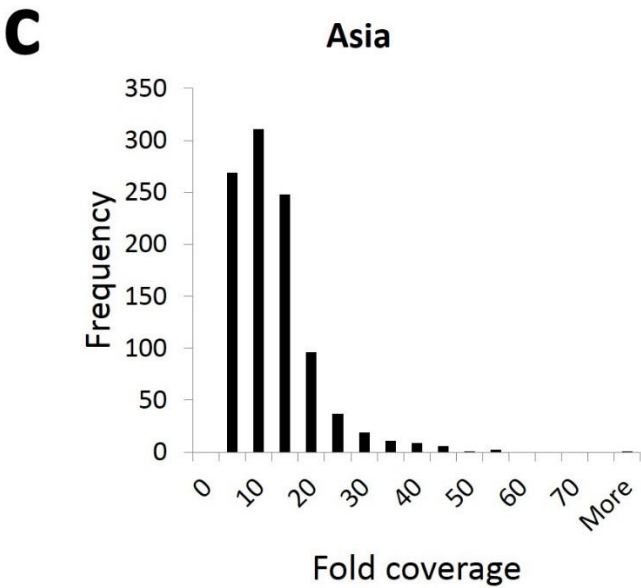
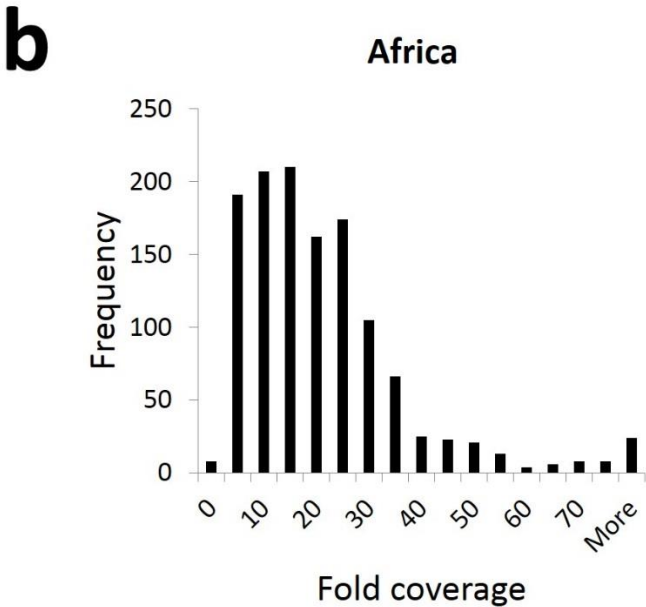
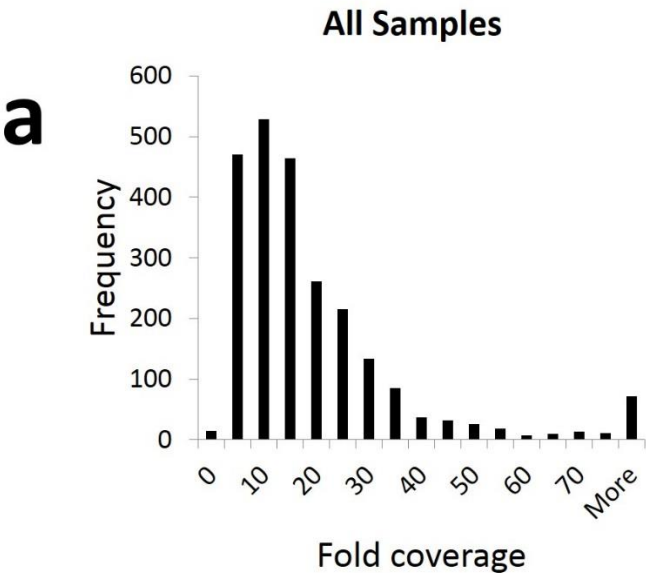
7.4 Sequences used in the *msp1b2*RefLib

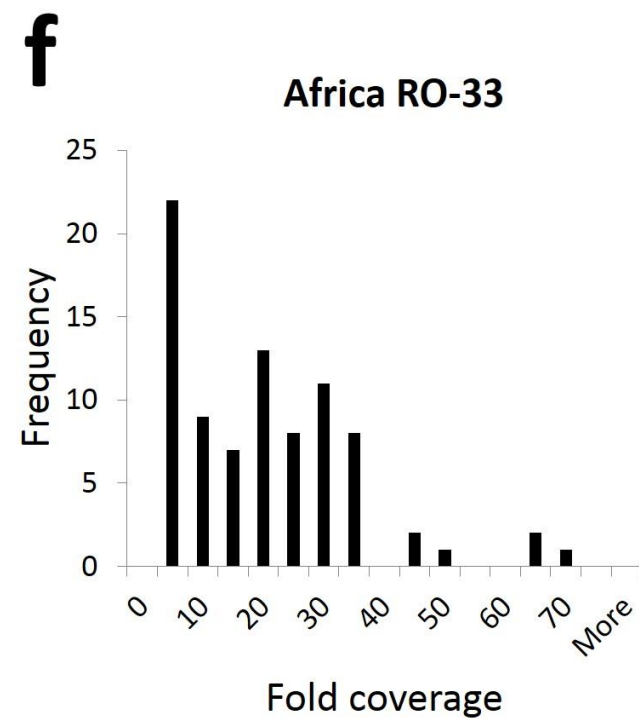
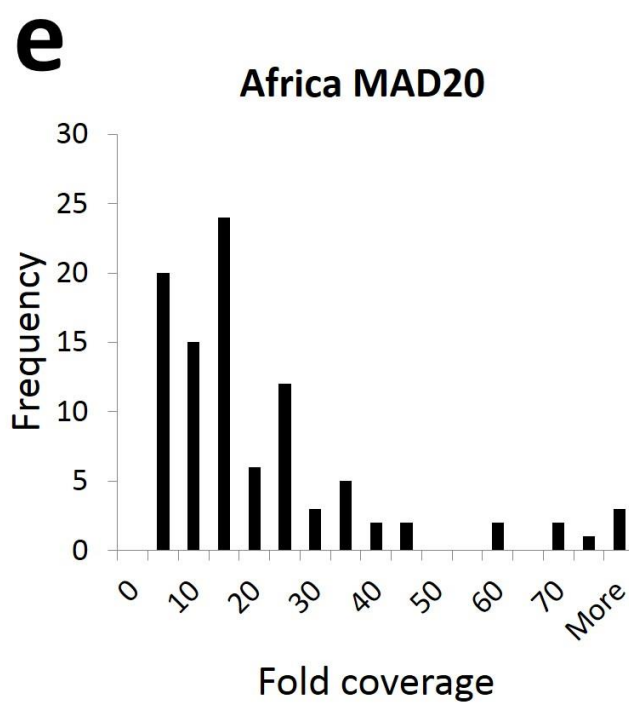
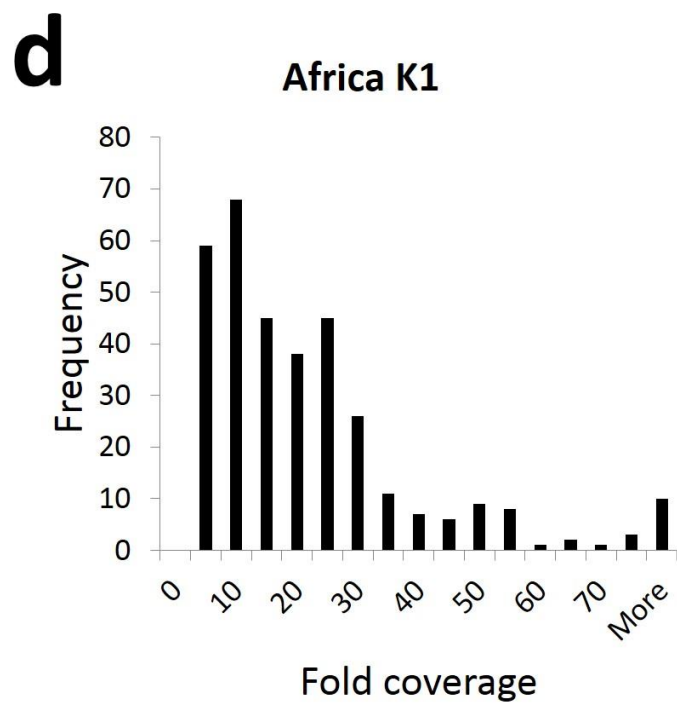
[illegible]

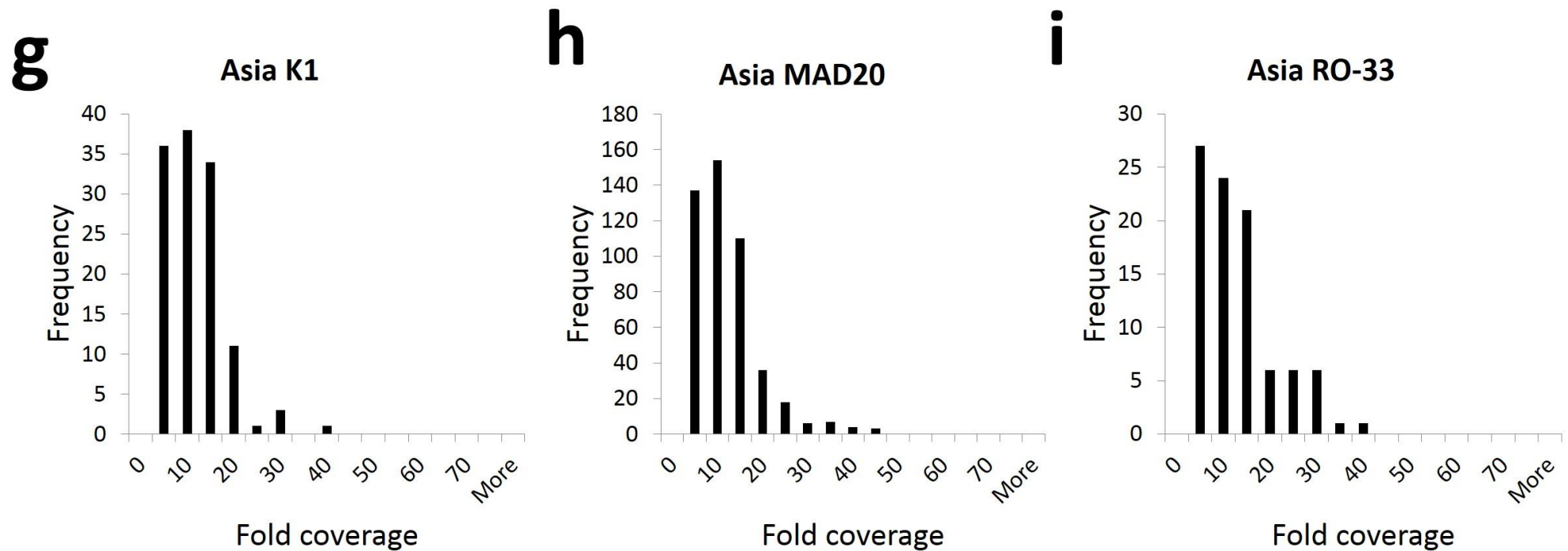
AB116598	aaaatggtattaAAGGATGGAGCAAATACTCAAGTTGTTGCAAAGCCTGCAGGTGCTGTAAGTACTCAAAGTGCTAAAAATCCTCCAGGTGCTACAGTACCTTCAG GTACTGCAAGTACTAAAGGTGCTATAAGATCTCCAGGTGCTGCAaatccttcagat
AB715486	aaaatggtattaAAGGATGGAGCAAATACTCAAGTTGTTGCAAAGCCTGCAGGTGCTGTAAGTACTCAAAGTGCTAAAAATCCTCCAGGTGCTACAGTACCTTCAG GTACTGCAAGTACTAAAGGTGCTATAAGATCTCCAGGTGCTGCAaatccttcagat
AB300614	aaaatggtattaAAGGATGGAGCAAATACTCAAGTTGTTGCAAAGCCTGCAGATGCTGTAAGTACTCAAAGTGCTAAAAATCCTCCAGGTGCTACAGTACCTTCAG GTACTGCAAGTACTAAAGGTGCTATAAGATCTCCAGGTGCTGCAaatccttcagat

The GenBank accession numbers of all sequences used in the *msp1b2*RefLib (section 2.3.4) are list grouped by allelic type. The *msp1* block 2 sequence is shown in capitals, with the block 1 and block 3 sequence fragments, also included in the library, shown in lower case.

7.5 Coverage of the *m*sp1b2RefLib by reads from Pf3k data







Short reads from the Pf3k project were aligned to the *msp1b2*RefLib (appendix 6.5). Fold coverage was calculated for each allelic type by dividing the number of bases in the reads mapped to sequences of a given allelic type by the total length of reference sequences of that allelic type. Coverage was then calculated for each of the 2400 samples by summing the coverage of each allelic type. The distribution of coverage across all samples (a) is shown (bin width = 5). Due to increased coverage of culture adapted samples (section 2.3.7) these were removed from further analysis. The distribution of coverage across (b) African samples was higher than (c) Asian samples ($p < 0.001$, Wilcoxon signed rank test). The coverage across (d) K1-like, (e) MAD20-like and (f) RO-33-like sequences from African samples with a single allelic type of *msp1* block 2 is not significantly varied ($p > 0.5$, Wilcoxon signed rank test). The coverage across (g) K1-like, (h) MAD20-like and (i) RO-33-like sequences from Asian samples with a single allelic type of *msp1* block 2 is not significantly varied ($p > 0.4$, Wilcoxon signed rank test).

7.6 Map showing location of sites of studies contributing to the Pf3k project

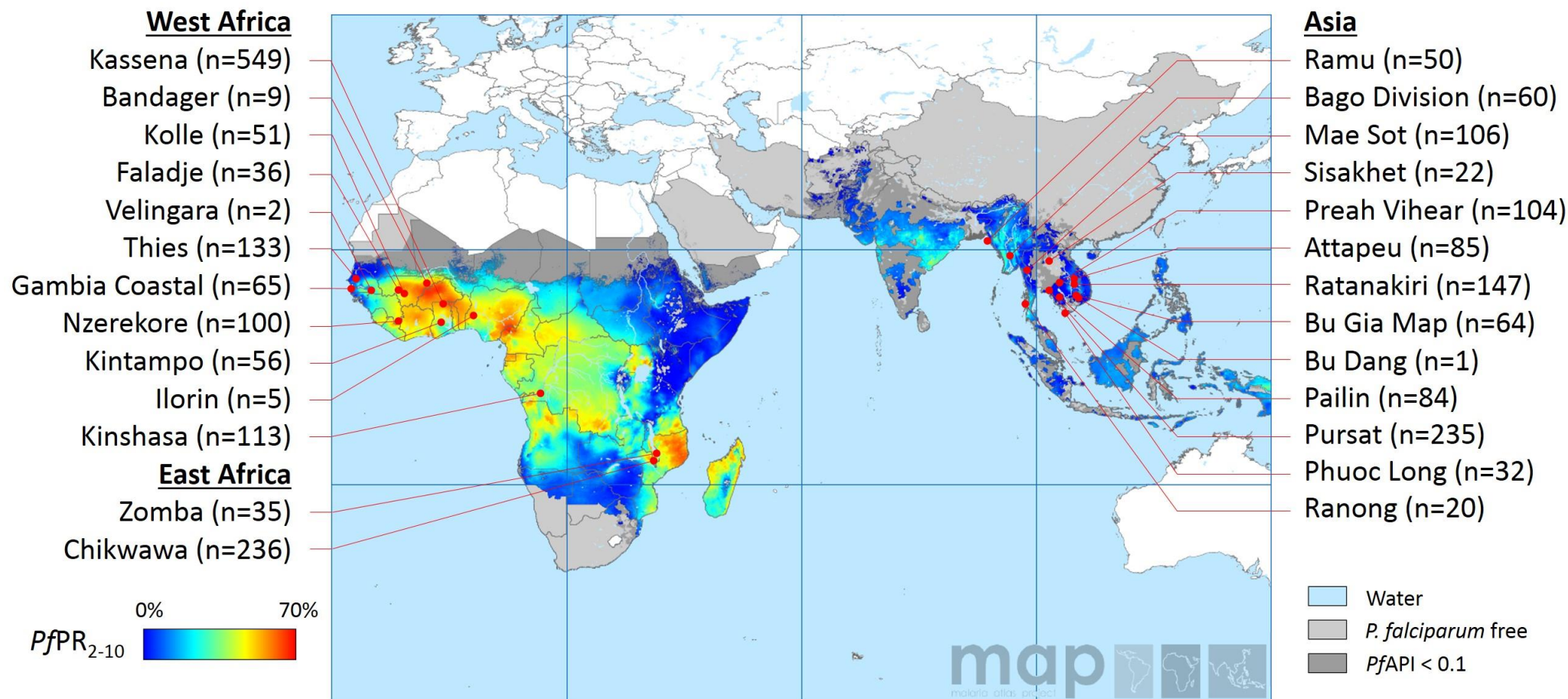


Figure 6.1 Pf3k study sites cover range of malaria endemicity. Map showing the location of the 25 sites from which samples in the Pf3k project were collected. All sites are shown as red dots and labelled with the name of the site with the number of samples in parentheses. The map, adapted from Gething *et al* 2011, is coloured according to endemicity of *P. falciparum* with a continuum from dark blue (0%) to red (70%) indicating the age-standardised annual mean percentage of children aged 2-10 years predicted to be positive for *P. falciparum* parasites (*PfPR*₂₋₁₀) in 2010 based on surveys of parasite prevalence conducted between January 1985 and June 2010. Countries with unstable transmission, defined as an annual incidence of *P. falciparum* (*PfAPI*) of less than 0.1 per 1000 people in 2010 or no transmission, defined as *PfAPI* of zero in 2010 are coloured dark and light grey, respectively. Countries not contributing data are coloured white.

7.7 Python functions for translating aligned reads, obtaining and analysing nonamers

The python functions for translating aligned reads (translate_reads), splitting translated sequences into nonamers (get_all_nonamers) and determining the frequency of nonamers (analyse_nonamers) are shown below.

```
import os
import re
import pandas
import pickle
import subprocess

ref_length={"K1_AB502485":327,
"K1_HM153224":147,
"K1_AB502454":309,
"K1_JX416340":183,
"K1_DQ485422":204,
"MAD20_AB502473":255,
"MAD20_AB502471":219,
"MAD20_Thai807":192,
"MAD20_DQ026702":156,
"MAD20_HM153185":159,
"RO-33_AB502489":162,
"RO-33_HQ821872":147,
"RO-33_B440":162,
"RO-33_DQ485450":156,
"RO-33_KP318438":162,}

codon_library = {'TTT':'F', 'TTC':'F', 'TTA':'L', 'TTG':'L','TCT':'S', 'TCC':'S', 'TCA':'S',
'TCG':'S',
'TAT':'Y', 'TAC':'Y', 'TAA':'*', 'TAG':'*', 'TGT':'C', 'TGC':'C', 'TGA':'*',
'TGG':'W', 'CTT':'L',
'CTC':'L', 'CTA':'L', 'CTG':'L', 'CCT':'P', 'CCC':'P', 'CCA':'P', 'CCG':'P', 'CAT':'H',
'CAC':'H',
'CAA':'Q', 'CAG':'Q', 'CGT':'R', 'CGC':'R', 'CGA':'R', 'CGG':'R', 'ATT':'I', 'ATC':'I',
'ATA':'I',
'ATG':'M', 'ACT':'T', 'ACC':'T', 'ACA':'T', 'ACG':'T', 'AAT':'N', 'AAC':'N', 'AAA':'K',
'AAG':'K',
'AGT':'S', 'AGC':'S', 'AGA':'R', 'AGG':'R', 'GTT':'V', 'GTC':'V', 'GTA':'V',
'GTG':'V', 'GCT':'A',
'GCC':'A', 'GCA':'A', 'GCG':'A', 'GAT':'D', 'GAC':'D', 'GAA':'E', 'GAG':'E', 'GGT':'G',
'GGC':'G',
'GGA':'G', 'GGG':'G', 'TCN':'S', 'TTN':'?', 'TAN':'?', 'TGN':'?', 'CTN':'L', 'CCN':'P',
'CAN':'?',
'CGN':'R', 'ATN':'?', 'ACN':'T', 'AAN':'?', 'AGN':'?', 'GTN':'V', 'GCN':'A', 'GAN':'?',
'GGN':'G',
'ANA':'?', 'ANT':'?', 'ANG':'?', 'ANC':'?', 'TNA':'?', 'TNT':'?', 'TNG':'?', 'TNC':'?', 'GNA':'?', 'GNT':
':?',
'GNG':'?', 'GNC':'?', 'CNA':'?', 'CNT':'?', 'CNG':'?', 'CNC':'?', 'NAA':'?', 'NAT':'?', 'NAG':'?', 'NAC':
':?',
'NTA':'?', 'NTT':'?', 'NTG':'?', 'NTC':'?', 'NGA':'?', 'NGT':'?', 'NGG':'?', 'NGC':'?', 'NCA':'?', 'NCT':
':?',
'NCG':'?', 'NCC':'?', 'NAN':'?',
'NTN':'?', 'NGN':'?', 'NCN':'?', 'NNA':'?', 'NNT':'?', 'NNG':'?', 'NNC':'?',
'ANN':'?', 'TNN':'?', 'GNN':'?', 'CNN':'?', 'NNN':'?' }
```

```
rp_mapping = {0:0,2:1,1:2}

def save_obj(obj, name ):
    with open(name + '.pkl', 'wb') as f:
        pickle.dump(obj, f, pickle.HIGHEST_PROTOCOL)

def load_obj(name ):
    with open(name + '.pkl', 'rb') as f:
        return pickle.load(f)

def build_allele_dictionary(csv_with_all_alleles):
    allele_dictionary = {}
```

```

with open(csv_with_all_alleles, "r") as i:
    lines = i.readlines()
for line in lines[1:]:
    sid = line.split(",")[1]
    allele = line.split(",")[2]
    allele_dictionary[sid] = allele
save_obj(allele_dictionary, "allele_dictionary")

def unpack_cigar(cigar_string):
#will find longest matching region and return dictionary with values #for M,I, D and S five
prime of first matching base of longest #matching region and three prime of this base
(including this base)
    cigar = {"M5":0,"I5":0,"D5":0,"S5":0,"M3":0,"I3":0,"D3":0,"S3":0}
    operations = filter(None, re.split("[0-9]+", cigar_string))
    values = filter(None, re.split("[M I D N S H P = X]+", cigar_string))
    op_vals = zip(values,operations)
    long_window = 0
    for o_v in op_vals:
        if o_v[1] == "M":
            match_window = int(o_v[0])
            if match_window > long_window:
                long_window = match_window
                best_match_o_v = o_v
    for i in op_vals[:op_vals.index(best_match_o_v)]:
        if i[1] == "M":
            cigar["M5"] = cigar["M5"]+int(i[0])
        else:
            if i[1] == "I":
                cigar["I5"] = cigar["I5"]+int(i[0])
            else:
                if i[1] == "D":
                    cigar["D5"] = cigar["D5"]+int(i[0])
                else:
                    if i[1] == "S":
                        cigar["S5"] = cigar["S5"]+int(i[0])
    for i in op_vals[op_vals.index(best_match_o_v):]:
        if i[1] == "M":
            cigar["M3"] = cigar["M3"]+int(i[0])
        else:
            if i[1] == "I":
                cigar["I3"] = cigar["I3"]+int(i[0])
            else:
                if i[1] == "D":
                    cigar["D3"] = cigar["D3"]+int(i[0])
                else:
                    if i[1] == "S":
                        cigar["S3"] = cigar["S3"]+int(i[0])
    return(cigar)

def find_frame(ref_start_pos, cigar):
    x = rp_mapping[ref_start_pos%3]

#determines number of bases that need to be clipped if read started #at best mapping position
    y = x + cigar["S5"] + cigar["M5"] + cigar["I5"]

#adds x to get total number bases prior to reference point
    z = y%3
    return(z)

def translate(read):
#input mapping read output region of read mapping to ref as list of #amino acids
    with open("Skipped_reads.csv", "w") as o:
        o.write("")
    ref = read.split("\t")[2]
    seq = read.split("\t")[9]
    ref_pos = int(read.split("\t")[3])-1
    cigar_string = read.split("\t")[5]
    if any(x in cigar_string for x in ["P","N","H","=", "X"]):

```

```

        with open("Skipped_reads.csv", "a") as o:
            o.write(read)
        pass
    cigar = unpack_cigar(cigar_string)
    ref_start_pos = ref_pos+cigar["M5"]+cigar["D5"]
    ref_end_pos = ref_pos+cigar["M5"]+cigar["M3"]+cigar["S3"]
    if ref_pos - cigar["S5"] < 0:
        seq_start_pos = cigar["S5"]-ref_pos
        cigar["S5"] = cigar["S5"]-seq_start_pos
    else:
        seq_start_pos = 0
    if ref_end_pos <= ref_length[ref]:
        seq_end_pos = len(seq)
    else:
        if ref_end_pos > ref_length[ref]:
            seq_end_pos = len(seq) - (ref_end_pos-ref_length[ref])
    f = find_frame(ref_start_pos, cigar)
    return("".join([codon_library[seq[i:i+3]] for i in range(seq_start_pos+f, ((seq_end_pos-
seq_start_pos-f)/3)*3)+(seq_start_pos-f),3)))

def translate_reads(bam, translated_read_dir):
    sample_id = str(bam).split("/")[-1]
    sam = sample_id+"_mapped.sam"
    translated_reads = []
    subprocess.call(["samtools", "view", "-h", "-o", sam, bam])
    with open(sam, "r") as sam:
        lines = sam.readlines()
    for read in lines:
        if read.startswith("@"):
            pass
        else:
            if read.split("\t")[2] != "*" and read.split("\t")[5] != "*":
#checks that read is mapped and has cigar string (i.e. is itself #mapped not just its mate
pair)
                transalted_read = translate(read)
                allele_type = read.split("\t")[2].split("_")[0]
                with open(translated_read_dir+"/"+sample_id+"_translated.csv", "a") as o:
                    o.write(read.split("\t")[0]+","+transalted_read+"\n")
                translated_reads.append((transalted_read, allele_type))
    sam = sample_id+"_mapped.sam"
    subprocess.call(["rm", str(sam)])
    return(translated_reads)

def nonamerise(amino_acid_sequence):
#input string of amino acid sequence output list of all nonamers
    return([amino_acid_sequence[i:i+9] for i in range(0, len(amino_acid_sequence)-9, 1)])

def get_all_nonamers(directory_with_all_bams):
#will take each sam in directory, translate all mapped reads and #output to new directory
#(..._translated_reads)
#will output csv file containing info on nonamers
    translated_read_dir = "Translated_reads"
    if not os.path.isdir(translated_read_dir):
        os.mkdir(translated_read_dir)
    all_nonamers = []
    with open("all_nonamers.csv", "w") as o:
        o.write("Sample_id,Fraction_of_total_reads_translated,Nonamer_sequence,Alle_type\n")
    for bam in os.listdir(directory_with_all_bams):
        if bam.endswith(".bam"):
            print(directory_with_all_bams+"/"+bam)
            translated_reads = translate_reads(directory_with_all_bams+"/"+bam,
translated_read_dir)
            for translated_read in translated_reads:
                nonamers = nonamerise(translated_read[0])
                for nonamer in nonamers:
                    with open("all_nonamers.csv", "a") as o:

```

```

        o.write(bam[:-
11]+","+str(float(1)/float(len(translated_reads)))+","+nonamer+', '+translated_read[1]+"\\n")
def analyse_nonamers(csv_with_all_nonamers):
    nonamers = pandas.read_csv(csv_with_all_nonamers)
    for un in set(nonamers["Nonamer_sequence"]):
        df = nonamers.loc[nonamers["Nonamer_sequence"] == un]
        weighted_frequency = sum(df["Fraction_of_total_reads_translated"])
        with open("nonamer_summary.csv", "a") as o:
            o.write(un+","+str(weighted_frequency)+"\\n")

```

7.8 Python script for algorithm to design polyvalent hybrid antigens

The python script (compile_nonamers) for generating polyvalent hybrid antigen designs (section 3.3.3) is shown below.

```
def compile_nonamers(nonamer_list, outfasta, maxlength, iterations, new_seed_propensity,
min_overlap):#nonamer_list needs to be csv sorted by frequency with nonamer seq in first
column
    nonamers = []
    length = 0
    seeds = []
    with open(nonamer_list, "r") as i:
        lines = i.readlines()
    for line in lines:
        nonamers.append((line.split(",") [0],float(line.split(",") [1].strip("\n"))))
    for x in range(0,iterations,1):
        nonamer_matches = []
        if length >= maxlength:#checks max length has not been exceded
            break
        else:
            n = 0
            for nonamer in nonamers:
                i=0
                overhangs = []
                for seed in seeds:
                    if nonamer[0] in seed:
                        nonamers.remove(nonamer)
                        break
                else:
                    left = find_left_overhang(seed, nonamer[0], min_overlap)
                    right = find_right_overhang(seed, nonamer[0], min_overlap)
                    if left == None and right == None:#no matching ends found
                        pass
                    else:
                        if left == None and right != None:#right match found
                            overhangs.append((right, "r", i))#triple denotes the overhang,
#right or left and the index of the seed
                        else:
                            if left != None and right == None:#left match found
                                overhangs.append((left, "l", i))
                            else:#overlaps at both ends so store best overlap (i.e.
#shortest overhang), NB right hand overhangs chosen if both equal
                                if len(right) <= len(left):
                                    overhangs.append((right, "r", i))
                                else:
                                    if len(left) < len(right):
                                        overhangs.append((left, "l", i))
                i+=1
                shortest = 9
                if len(overhangs) == 0:#no end matches found nonamer stored as a potential new
#seed
                    nonamer_matches.append((nonamer[0], "n", None, len(nonamer[0])-
new_seed_propensity*nonamer[1],n))#creates quintuple with nonamer/overhang sequence; left of
seed, right of seed or new seed; seed index; inclusion score; nonamer index)
                else:
                    for overhang in overhangs:
                        if len(overhang[0]) < shortest:
                            shortest = overhang[0]
                            best_overhang = overhang
                    nonamer_matches.append(best_overhang+(len(best_overhang[0])-
new_seed_propensity*nonamer[1],n))#creates quintuple with nonamer/overhang sequence; left of
seed, right of seed or new seed; seed index; inclusion score; nonamer index)
                    n+=1
                    include = 9
                    for nonamer_match in nonamer_matches:
                        if nonamer_match[3] < include:
                            include = nonamer_match[3]
                            nonamer_to_include = nonamer_match
                    nonamers =
nonamers[:nonamer_to_include[4]]+nonamers[nonamer_to_include[4]+1:]#removes nonamer from
#list
                    if nonamer_to_include[1] == "r":
                        seeds[nonamer_to_include[2]] =
seeds[nonamer_to_include[2]]+nonamer_to_include[0]
                    else:
                        if nonamer_to_include[1] == "l":
```



```

        seeds[nonamer_to_include[2]] =
nonamer_to_include[0]+seeds[nonamer_to_include[2]]
    else:
        if nonamer_to_include[1] == "n":
            seeds.append(nonamer_to_include[0])
        seeds = merge_seeds(seeds, min_overlap)
        length = sum(len(s) for s in seeds)
        print(x)
        print(seeds)
    antigen = "-".join(seeds)
    print(antigen)
    with open(outfasta, "w") as o:
        o.write("> "+nonamer_list+"_antigen_maxlength_"+str(maxlength)+"\n"+antigen)

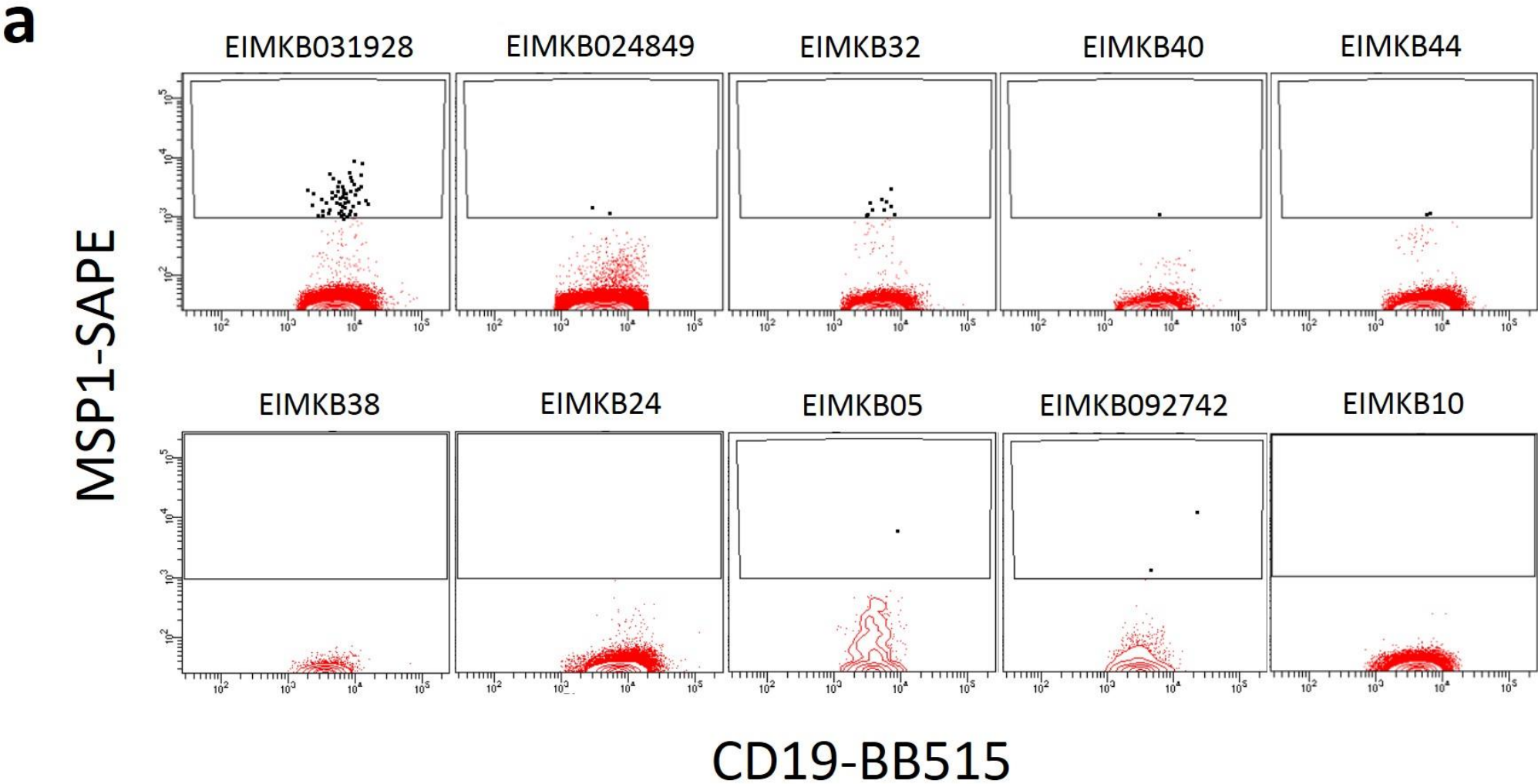
def find_right_overhang(seed, nonamer, min_overlap):
    for i in range(len(nonamer)-1,0+(min_overlap-1),-1):
        if nonamer[i:] == seed[-i:]:
            return(nonamer[i:])

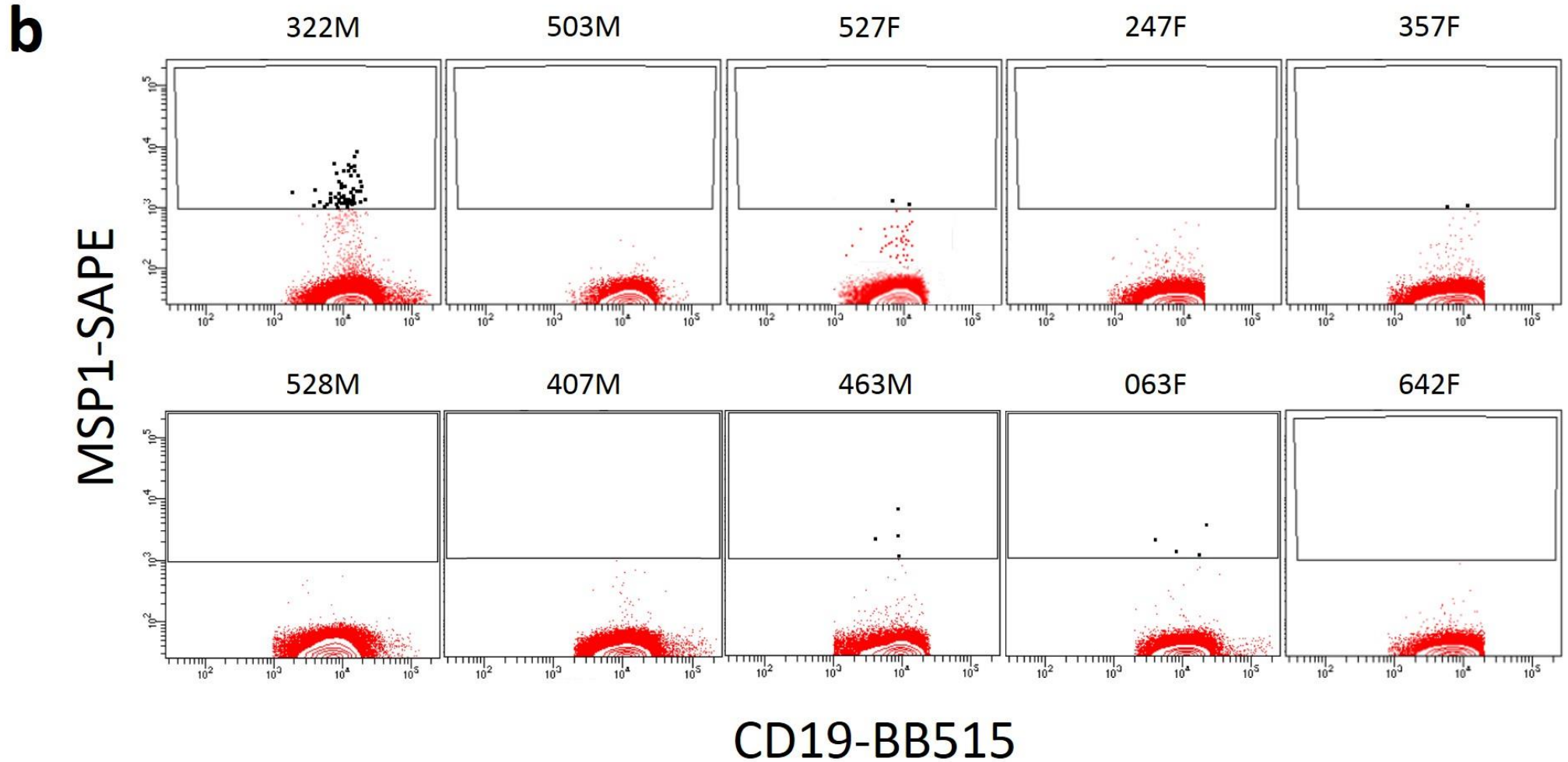
def find_left_overhang(seed, nonamer, min_overlap):
    for i in range(1,len(nonamer)+1-min_overlap,1):
        if nonamer[i:] == seed[:9-i]:
            return(nonamer[i:])

def merge_seeds(seeds, min_overlap):
    for seed in seeds:
        seeds.remove(seed)
        if any(seed in string for string in seeds):
            pass
        else:
            seeds.append(seed)
    matches = []
    combinations = list(itertools.combinations(seeds,2))
    for combo in combinations:
        left = find_left_overhang(combo[0], combo[1], min_overlap)
        right = find_right_overhang(combo[0], combo[1], min_overlap)
        if left == None and right == None:#no matching ends found
            pass
        else:
            if left == None and right != None:#right match found
                matches.append((right, "r", combo))#triple denotes the overhang, right or
#left and the index of the seed
            else:
                if left != None and right == None:#left match found
                    matches.append((left, "l", combo))
                else:#overlaps at both ends so store best overlap (i.e. shortest overhang), NB
#right hand overhangs chosen if both equal
                    if len(right) <= len(left):
                        matches.append((right, "r", combo))
                    else:
                        if len(left) < len(right):
                            matches.append((left, "l", combo))
    for match in matches:
        if match[1] == "r":
            seeds.append(match[2][0]+match[0])
        else:
            if match[1] == "l":
                seeds.append(match[0]+match[2][0])
            seeds.remove(match[2][0])
            seeds.remove(match[2][1])
    return(seeds)

```

7.9 FACS plots showing labelling of memory B-cells with MSP-1-SAPE antigen tetramers





B-cells enriched from peripheral blood mononuclear cells (PBMC) taken from malaria exposed (a) or naïve (b) donors were labelled with FVS-780 anti-CD19-BB515, anti-CD27-APC, anti-CD3-PerCP, anti-CD14-PerCP, anti-CD16-PerCP and MSP-1-SAPE tetramers (section 4.2.5). Dead (FVS-780 positive) and non-B (CD3/CD14/CD16 positive) were gated out prior to gating of CD19, CD27 double positive memory B-cells. Plots show fluorescence (arbitrary units) in the BB515 channel against fluorescence in the PE channel. Cells above 10^3 fluorescence units in the PE channel (black dots) were sorted as MSP-1-SAPE positive cells. MSP-1-SAPE negative cells (red contour plot) were sent to waste.

7.10 List of additional data files

File name	Description
long_read_sequences.fa	Fasta file containing all sequences in LRD
Pf3k_short_read_assembled_translated_sequences.csv	CSV file containing all amino acid sequences encoded by sequences assembled from Pf3k short read data. Column 1: Sample ID, column 2: Region, column 3: country, column 4: site, column 5: clone (for samples yielding multiple sequences, denoted as c1...c4), column 6: allelic type, column 7: length of MSP-1 block 2 length (number of amino acid residues), column 8: predicted MSP-1 block 2 amino acid sequence.

Table shows all additional data files with a description of their contents. All files can be found on attached compact disc or downloaded from <https://tinyurl.com/HAJ2017>. Abbreviations: ID-identification, CSV - comma separated values.

7.11 References

- AHMEDOU SALEM, M. S., NDIAYE, M., OULDABDALLAHI, M., LEKWEIRY, K. M., BOGREAU, H., KONATE, L., FAYE, B., GAYE, O., FAYE, O. & MOHAMED SALEM, O. B. A. O. 2014. Polymorphism of the merozoite surface protein-1 block 2 region in *Plasmodium falciparum* isolates from Mauritania. *Malar J*, 13, 26.
- AL-ABD, N. M., MAHDY, M. A., AL-MEKHLAFI, A. M., SNOUNOU, G., ABDUL-MAJID, N. B., AL-MEKHLAFI, H. M. & FONG, M. Y. 2013. The suitability of *P. falciparum* merozoite surface proteins 1 and 2 as genetic markers for in vivo drug trials in Yemen. *PLoS One*, 8, e67853.
- ALFADHLI, S. & ORJIH, A. 2004. Sick cell alters *Plasmodium falciparum* genomic DNA. *Unpublished*.
- APINJOH, T. O., TATA, R. B., ANCHANG-KIMBI, J. K., CHI, H. F., FON, E. M., MUGRI, R. N., TANGO, D. A., NYINGCHU, R. V., GHOGOMU, S. M., NKUO-AKENJI, T. & ACHIDI, E. A. 2015. *Plasmodium falciparum* merozoite surface protein 1 block 2 gene polymorphism in field isolates along the slope of Mount Cameroon: a cross-sectional study. *BMC Infect Dis*, 15, 309.
- AUBOUY, A., MIGOT-NABIAS, F. & DELORON, P. 2003. Polymorphism in two merozoite surface proteins of *Plasmodium falciparum* isolates from Gabon. *Malar J*, 2, 12.
- AYANFUL-TORGBY, R., OPPONG, A., ABANKWA, J., ACQUAH, F., WILLIAMSON, K. C. & AMOAH, L. E. 2016. *Plasmodium falciparum* genotype and gametocyte prevalence in children with uncomplicated malaria in coastal Ghana. *Malar J*, 15, 592.
- BAMIDELE ABIODUN, I., OLUWADUN, A., OLUGBENGA AYOOLA, A. & SENAPON OLUOLA, I. 2016. *Plasmodium falciparum* Merozoite Surface Protein-1 Polymorphisms among Asymptomatic Sick Cell Anemia Patients in Nigeria. *Acta Med Iran*, 54, 44-53.
- BHARTI, P. K., SHUKLA, M., GAUTAM, S. P., SHARMA, Y. D., SINGH, N. 2011. Genetic complexity of merozoite surface protein 1 from central India. *Unpublished*.
- BOUYOU-AKOTET, M. K., M'BONDOKWE, N. P. & MAWILI-MBOUMBA, D. P. 2015. Genetic polymorphism of merozoite surface protein-1 in *Plasmodium falciparum* isolates from patients with mild to severe malaria in Libreville, Gabon. *Parasite*, 22, 12.
- BRANCH, O. H., TAKALA, S., KARIUKI, S., NAHLEN, B. L., KOLCZAK, M., HAWLEY, W. & LAL, A. A. 2001. *Plasmodium falciparum* genotypes, low complexity of infection, and resistance to subsequent malaria in participants in the Asembo Bay Cohort Project. *Infect Immun*, 69, 7783-92.
- CERTA, U., ROTMANN, D., MATILE, H. & REBER-LISKE, R. 1987. A naturally occurring gene encoding the major surface antigen precursor p190 of *Plasmodium falciparum* lacks tripeptide repeats. *EMBO J*, 6, 4137-42.
- CHANG, S. P., KRAMER, K. J., YAMAGA, K. M., KATO, A., CASE, S. E. & SIDDIQUI, W. A. 1988. *Plasmodium falciparum*: gene structure and hydropathy profile of the major merozoite surface antigen (gp195) of the Uganda-Palo Alto isolate. *Exp Parasitol*, 67, 1-11.
- COLBORN, J. M., BYRD, B. D., KOITA, O. A., KROGSTAD, D. J. 2006. Effects of adenine/thymine content and amplicon length on estimates of starting copy number using real-time PCR with SYBR green. *Unpublished*.
- CONWAY, D. J., CAVANAGH, D. R., TANABE, K., ROPER, C., MIKES, Z. S., SAKIHAMA, N., BOJANG, K. A., ODUOLA, A. M., KREMSNER, P. G., ARNOT, D. E., GREENWOOD, B. M. & MCBRIDE, J. S. 2000. A principal target of human immunity to malaria identified by molecular population genetic and immunological analyses. *Nat Med*, 6, 689-92.
- DOLMAZON, V., MATSIKA-CLAQUIN, M. D., MANIRAKIZA, A., YAPOU, F., NAMBOT, M. & MENARD, D. 2008. Genetic diversity and genotype multiplicity of *Plasmodium falciparum* infections in symptomatic individuals living in Bangui (CAR). *Acta Trop*, 107, 37-42.
- GETHING, P. W., PATIL, A. P., SMITH, D. L., GUERRA, C. A., ELYAZAR, I. R., JOHNSTON, G. L., TATEM, A. J. & HAY, S. I. 2011. A new world malaria map: *Plasmodium falciparum* endemicity in 2010. *Malar J*, 10, 378.

- GHANCHI, N. K., MARTENSSON, A., URSING, J., JAFRI, S., BEREZKY, S., HUSSAIN, R. & BEG, M. A. 2010. Genetic diversity among *Plasmodium falciparum* field isolates in Pakistan measured with PCR genotyping of the merozoite surface protein 1 and 2. *Malar J*, 9, 1.
- GOSI, P., LANTERI, C. A., TYNER, S. D., SE, Y., LON, C., SPRING, M., CHAR, M., SEA, D., SRIWICHA, S., SURASRI, S., WONGARUNKOCHAKORN, S., PIDTANA, K., WALSH, D. S., FUKUDA, M. M., MANNING, J., SAUNDERS, D. L. & BETHELL, D. 2013. Evaluation of parasite subpopulations and genetic diversity of the msp1, msp2 and glurp genes during and following artesunate monotherapy treatment of *Plasmodium falciparum* malaria in Western Cambodia. *Malar J*, 12, 403.
- HAMID, M. M., MOHAMMED, S. B. & EL HASSAN, I. M. 2013. Genetic Diversity of *Plasmodium falciparum* Field Isolates in Central Sudan Inferred by PCR Genotyping of Merozoite Surface Protein 1 and 2. *N Am J Med Sci*, 5, 95-101.
- HOLDER, A. A., LOCKYER, M. J., ODINK, K. G., SANDHU, J. S., RIVEROS-MORENO, V., NICHOLLS, S. C., HILLMAN, Y., DAVEY, L. S., TIZARD, M. L., SCHWARZ, R. T. & ET AL. 1985. Primary structure of the precursor to the three major surface antigens of *Plasmodium falciparum* merozoites. *Nature*, 317, 270-3.
- JIANG, G., DAUBENBERGER, C., HUBER, W., MATILE, H., TANNER, M. & PLUSCHKE, G. 2000. Sequence diversity of the merozoite surface protein 1 of *Plasmodium falciparum* in clinical isolates from the Kilombero District, Tanzania. *Acta Trop*, 74, 51-61.
- JIANG, G., LIU, R., DAUBENBERGER, C. A. & PLUSCHKE, G. 1999. Sequence analysis of the MSP 1 gene of *Plasmodium falciparum* isolates from Hainan, China. *Zhongguo Ji Sheng Chong Xue Yu Ji Sheng Chong Bing Za Zhi*, 17, 294-7.
- JONGWUTIWES, S., TANABE, K., NAKAZAWA, S., YANAGI, T. & KANBARA, H. 1992. Sequence variation in the tripeptide repeats and T cell epitopes in P190 (MSA-1) of *Plasmodium falciparum* from field isolates. *Mol Biochem Parasitol*, 51, 81-9.
- JOSHI, H., VALECHA, N., VERMA, A., KAUL, A., MALLICK, P. K., SHALINI, S., PRAJAPATI, S. K., SHARMA, S. K., DEV, V., BISWAS, S., NANDA, N., MALHOTRA, M. S., SUBBARAO, S. K. & DASH, A. P. 2007. Genetic structure of *Plasmodium falciparum* field isolates in eastern and north-eastern India. *Malar J*, 6, 60.
- JOSHI, H., VERMA, A. & DAS, M. K. 2005. Malaria among Jarawas-the primitive and hostile tribe of Andaman and Nicobar Islands India. *Unpublished*.
- JULIANO, J. J., PORTER, K., MWAPASA, V., SEM, R., ROGERS, W. O., ARIEY, F., WONGSRICHANALAI, C., READ, A. & MESHNICK, S. R. 2010. Exposing malaria in-host diversity and estimating population diversity by capture-recapture using massively parallel pyrosequencing. *Proc Natl Acad Sci U S A*, 107, 20138-43.
- KANG, J. M., MOON, S. U., KIM, J. Y., CHO, S. H., LIN, K., SOHN, W. M., KIM, T. S. & NA, B. K. 2010. Genetic polymorphism of merozoite surface protein-1 and merozoite surface protein-2 in *Plasmodium falciparum* field isolates from Myanmar. *Malar J*, 9, 131.
- KHAMINSOU, N., KRITPETCHARAT, O., DADUANG, J., CHARERNTANYARAK, L. & KRITPETCHARAT, P. 2011. Genetic analysis of the merozoite surface protein-1 block 2 allelic types in *Plasmodium falciparum* clinical isolates from Lao PDR. *Malar J*, 10, 371.
- KIMURA, E., MATTEI, D., DI SANTI, S. M. & SCHERF, A. 1990. Genetic diversity in the major merozoite surface antigen of *Plasmodium falciparum*: high prevalence of a third polymorphic form detected in strains derived from malaria patients. *Gene*, 91, 57-62.
- KOLAWOLE, O. M., MOKUOLU, O. A., OLUKOSI, Y. A. & OLOYEDE, T. O. 2016. Population genomics diversity of *Plasmodium falciparum* in malaria patients attending Okelele Health Centre, Okelele, Ilorin, Kwara State, Nigeria. *Afr Health Sci*, 16, 704-711.
- KWIEK, J. J., ALKER, A. P., WENINK, E. C., CHAPONDA, M., KALILANI, L. V. & MESHNICK, S. R. 2007. Estimating true antimalarial efficacy by heteroduplex tracking assay in patients with complex *Plasmodium falciparum* infections. *Antimicrob Agents Chemother*, 51, 521-7.

- LILJANDER, A., WIKLUND, L., FALK, N., KWEKU, M., MARTENSSON, A., FELGER, I. & FARNERT, A. 2009. Optimization and validation of multi-coloured capillary electrophoresis for genotyping of *Plasmodium falciparum* merozoite surface proteins (msp1 and 2). *Malar J*, 8, 78.
- MACKAY, M., GOMAN, M., BONE, N., HYDE, J. E., SCAIFE, J., CERTA, U., STUNNENBERG, H. & BUJARD, H. 1985. Polymorphism of the precursor for the major surface antigens of *Plasmodium falciparum* merozoites: studies at the genetic level. *EMBO J*, 4, 3823-9.
- MAHDI ABDEL HAMID, M., ELAMIN, A. F., ALBSHEER, M. M., ABDALLA, A. A., MAHGOUB, N. S., MUSTAFA, S. O., MUNEEER, M. S. & AMIN, M. 2016. Multiplicity of infection and genetic diversity of *Plasmodium falciparum* isolates from patients with uncomplicated and severe malaria in Gezira State, Sudan. *Parasit Vectors*, 9, 362.
- MAWILI-MBOUMBA, D. P., MBONDOKWE, N., ADANDE, E. & BOUYOU-AKOTET, M. K. 2015. Allelic Diversity of MSP1 Gene in *Plasmodium falciparum* from Rural and Urban Areas of Gabon. *Korean J Parasitol*, 53, 413-9.
- MAYENGUE, P. I., NDOUNGA, M., MALONGA, F. V., BITEMO, M. & NTOUMI, F. 2011. Genetic polymorphism of merozoite surface protein-1 and merozoite surface protein-2 in *Plasmodium falciparum* isolates from Brazzaville, Republic of Congo. *Malar J*, 10, 276.
- MEDEIROS, M. M., FOTORAN, W. L., DALLA MARTHA, R. C., KATSURAGAWA, T. H., PEREIRA DA SILVA, L. H. & WUNDERLICH, G. 2013. Natural antibody response to *Plasmodium falciparum* merozoite antigens MSP5, MSP9 and EBA175 is associated to clinical protection in the Brazilian Amazon. *BMC Infect Dis*, 13, 608.
- MOBASSIR, H. M., SOHAIL, M., RITESH, K., ORALEE, B. H., TRIDIBES, A., RAZIUDDIN, M. 2010. Genetic Diversity in Merozoite Surface Protein-1 and 2 among *Plasmodium falciparum* Isolates from Malarious Districts of Tribal Dominant State Jharkhand of India. *Unpublished*.
- MOHAMMED, H., MINDAYE, T., BELAYNEH, M., KASSA, M., ASSEFA, A., TADESSE, M., WOYESSA, A., MENGESHA, T. & KEBEDE, A. 2015. Genetic diversity of *Plasmodium falciparum* isolates based on MSP-1 and MSP-2 genes from Kolla-Shele area, Arbaminch Zuria District, southwest Ethiopia. *Malar J*, 14, 73.
- MOHD ABD RAZAK, M. R., SASTU, U. R., NORAHMAD, N. A., ABDUL-KARIM, A., MUHAMMAD, A., MUNIANDY, P. K., JELIP, J., RUNDI, C., IMWONG, M., MUDIN, R. N. & ABDULLAH, N. R. 2016. Genetic Diversity of *Plasmodium falciparum* Populations in Malaria Declining Areas of Sabah, East Malaysia. *PLoS One*, 11, e0152415.
- MWINGIRA, F., NKWENGULILA, G., SCHOEPLIN, S., SUMARI, D., BECK, H. P., SNOUNOU, G., FELGER, I., OLLIARO, P. & MUGITTU, K. 2011. *Plasmodium falciparum* msp1, msp2 and glurp allele frequency and diversity in sub-Saharan Africa. *Malar J*, 10, 79.
- MYLER, P. J. 1989. Nucleotide and deduced amino acid sequence of the gp195 (MSA-1) gene from *Plasmodium falciparum* Palo Alto PLF-3/B11. *Nucleic Acids Res*, 17, 5401.
- NIANG, M., LOUCOUBAR, C., SOW, A., DIAGNE, M. M., FAYE, O., FAYE, O., DIALLO, M., TOURE-BALDE, A. & SALL, A. A. 2016. Genetic diversity of *Plasmodium falciparum* isolates from concurrent malaria and arbovirus co-infections in Kedougou, southeastern Senegal. *Malar J*, 15, 155.
- NIANG, M., THIAM, L. G., LOUCOUBAR, C., SOW, A., SADIO, B. D., DIALLO, M., SALL, A. A. & TOURE-BALDE, A. 2017. Spatio-temporal analysis of the genetic diversity and complexity of *Plasmodium falciparum* infections in Kedougou, southeastern Senegal. *Parasit Vectors*, 10, 33.
- NORANATE, N., PRUGNOLLE, F., JOUIN, H., TALL, A., MARRAMA, L., SOKHNA, C., EKALA, M. T., GUILLOTTE, M., BISCHOFF, E., BOUCHIER, C., PATARAPOTIKUL, J., OHASHI, J., TRAPE, J. F., ROGIER, C. & MERCEREAU-PUIJALON, O. 2009. Population diversity and antibody selective pressure to *Plasmodium falciparum* MSP1 block2 locus in an African malaria-endemic setting. *BMC Microbiol*, 9, 219.
- OGOYEMI-HOUNTO, A., GAZARD, D. K., NDAM, N., TOPANOU, E., GARBA, O., ELEGBE, P., HOUNTOHOTEGBE, T. & MASSOUGBODJI, A. 2013a. Genetic polymorphism of merozoite

- surface protein-1 and merozoite surface protein-2 in *Plasmodium falciparum* isolates from children in South of Benin. *Parasite*, 20, 37.
- OGOUEMI-HOUNTO, A., NDAM, N. T., FADEGNON, G., AZAGNANDJI, C., BELLO, M., MOUSSILIOU, A., CHIPPAUX, J. P., KINDE GAZARD, D. & MASSOUGBODJI, A. 2013b. Low prevalence of the molecular markers of *Plasmodium falciparum* resistance to chloroquine and sulphadoxine/pyrimethamine in asymptomatic children in Northern Benin. *Malar J*, 12, 413.
- OLAFSSON, P., MATILE, H. & CERTA, U. 1992. *Plasmodium falciparum*: the repetitive MSA-1 surface protein of the RO-71 isolate is recognized by mouse antibody against the nonrepetitive repeat block of RO-33. *Exp Parasitol*, 74, 381-9.
- PAN, W., TOLLE, R. & BUJARD, H. 1995. A direct and rapid sequencing strategy for the *Plasmodium falciparum* antigen gene gp190/MSA1. *Mol Biochem Parasitol*, 73, 241-4.
- PETERSON, M. G., COPPEL, R. L., MCINTYRE, P., LANGFORD, C. J., WOODROW, G., BROWN, G. V., ANDERS, R. F. & KEMP, D. J. 1988. Variation in the precursor to the major merozoite surface antigens of *Plasmodium falciparum*. *Mol Biochem Parasitol*, 27, 291-301.
- PEYERL-HOFFMANN, G., JELINEK, T., KILIAN, A., KABAGAMBE, G., METZGER, W. G. & VON SONNENBURG, F. 2001. Genetic diversity of *Plasmodium falciparum* and its relationship to parasite density in an area with different malaria endemicities in West Uganda. *Trop Med Int Health*, 6, 607-13.
- RAJ, D. K., DAS, B. R., DASH, A. P. & SUPAKAR, P. C. 2004. Genetic diversity in the merozoite surface protein 1 gene of *Plasmodium falciparum* in different malaria-endemic localities. *Am J Trop Med Hyg*, 71, 285-9.
- RANFORD-CARTWRIGHT, L. C., BALFE, P., CARTER, R. & WALLIKER, D. 1991. Direct sequencing of enzymatically amplified DNA of alleles of the merozoite surface antigen MSA-1 gene from the malaria parasite *Plasmodium falciparum*. *Mol Biochem Parasitol*, 46, 185-7.
- ROBINSON, T., CAMPINO, S. G., AUBURN, S., ASSEFA, S. A., POLLEY, S. D., MANSKE, M., MACINNIS, B., ROCKETT, K. A., MASLEN, G. L., SANDERS, M., QUAIL, M. A., CHIODINI, P. L., KWIATKOWSKI, D. P., CLARK, T. G. & SUTHERLAND, C. J. 2011. Drug-resistant genotypes and multi-clonality in *Plasmodium falciparum* analysed by direct genome sequencing from peripheral blood of malaria patients. *PLoS One*, 6, e23204.
- SAHA, P., GANGULY, S. & MAJI, A. K. 2016. Genetic diversity and multiplicity of infection of *Plasmodium falciparum* isolates from Kolkata, West Bengal, India. *Infect Genet Evol*, 43, 239-44.
- SAHA, P., GUHA, S. K., DAS, S., MULLICK, S., GANGULY, S., BISWAS, A., BERA, D. K., CHATTOPADHYAY, G., DAS, M., KUNDU, P. K., RAY, K. & MAJI, A. K. 2012. Comparative efficacies of artemisinin combination therapies in *Plasmodium falciparum* malaria and polymorphism of pfATPase6, pfCRT, pfDHFR, and pfDhps genes in tea gardens of Jalpaiguri District, India. *Antimicrob Agents Chemother*, 56, 2511-7.
- SAKIHAMA, N., KIMURA, M., HIRAYAMA, K., KANDA, T., NA-BANGCHANG, K., JONGWUTIWES, S., CONWAY, D. & TANABE, K. 1999. Allelic recombination and linkage disequilibrium within Msp-1 of *Plasmodium falciparum*, the malignant human malaria parasite. *Gene*, 230, 47-54.
- SAKIHAMA, N., NAKAMURA, M., PALANCA, A. A., JR., ARGUBANO, R. A., REALON, E. P., LARRACAS, A. L., ESPINA, R. L. & TANABE, K. 2007. Allelic diversity in the merozoite surface protein 1 gene of *Plasmodium falciparum* on Palawan Island, the Philippines. *Parasitol Int*, 56, 185-94.
- SAKIHAMA, N., OHMAE, H., BAKOTE'E, B., KAWABATA, M., HIRAYAMA, K. & TANABE, K. 2006. Limited allelic diversity of *Plasmodium falciparum* merozoite surface protein 1 gene from populations in the Solomon Islands. *Am J Trop Med Hyg*, 74, 31-40.
- SCHERF, A., MATTEI, D. & SARTHOU, J. L. 1991. Multiple infections and unusual distribution of block 2 of the MSA1 gene of *Plasmodium falciparum* detected in west African clinical isolates by polymerase chain reaction analysis. *Mol Biochem Parasitol*, 44, 297-9.
- SCHLEIERMACHER, D., ROGIER, C., SPIEGEL, A., TALL, A., TRAPE, J. F. & MERCEREAU-PUJALON, O. 2001. Increased multiplicity of *Plasmodium falciparum* infections and skewed distribution of

- individual msp1 and msp2 alleles during pregnancy in Ndiop, a Senegalese village with seasonal, mesoendemic malaria. *Am J Trop Med Hyg*, 64, 303-9.
- SCOPEL, K. K., FONTES, C. J., FERREIRA, M. U. & BRAGA, E. M. 2005. Plasmodium falciparum: IgG subclass antibody response to merozoite surface protein-1 among Amazonian gold miners, in relation to infection status and disease expression. *Exp Parasitol*, 109, 124-34.
- SEHGAL, R., KAUR, H., BANSAL, D., SULTAN, A. A. 2015. Genetic diversity of Plasmodium falciparum Merozoite Surface Protein 1 (block2 region) from Chandigarh and adjoining states of North India. *Unpublished*.
- SEHGAL, R., KAUR, H., SINGH, V. P., GOYAL, K., BANSAL, D., SULTAN, A. A. 2014. Study the molecular epidemiology and genetic diversity of malaria parasites in India and in Qatar. *Unpublished*.
- SHAN, Z. X., YU, X. B., LI, X. R., MA, C. L., FANG, J. M. 1999. Molecular cloning and sequence analysis of major merozoite surface antigen(gp195) gene of Plasmodium falciparum isolate FCC1/HN *Unpublished*.
- SOULAMA, I., NEBIE, I., OUEDRAOGO, A., GANSANE, A., DIARRA, A., TIONO, A. B., BOUGOUMA, E. C., KONATE, A. T., KABRE, G. B., TAYLOR, W. R. & SIRIMA, S. B. 2009. Plasmodium falciparum genotypes diversity in symptomatic malaria of children living in an urban and a rural setting in Burkina Faso. *Malar J*, 8, 135.
- SULISTYANINGSIH, E., FITRI, L. E., LOSCHER, T. & BERENS-RIHA, N. 2013. Diversity of the var gene family of Indonesian Plasmodium falciparum isolates. *Malar J*, 12, 80.
- TAKALA, S., BRANCH, O., ESCALANTE, A. A., KARIUKI, S., WOOTTON, J. & LAL, A. A. 2002. Evidence for intragenic recombination in Plasmodium falciparum: identification of a novel allele family in block 2 of merozoite surface protein-1: Asembo Bay Area Cohort Project XIV. *Mol Biochem Parasitol*, 125, 163-71.
- TAKALA, S. L., ESCALANTE, A. A., BRANCH, O. H., KARIUKI, S., BISWAS, S., CHAIYAROJ, S. C. & LAL, A. A. 2006. Genetic diversity in the Block 2 region of the merozoite surface protein 1 (MSP-1) of Plasmodium falciparum: additional complexity and selection and convergence in fragment size polymorphism. *Infect Genet Evol*, 6, 417-24.
- TANABE, K., MACKAY, M., GOMAN, M. & SCAIFE, J. G. 1987. Allelic dimorphism in a surface antigen gene of the malaria parasite Plasmodium falciparum. *Journal of Molecular Biology*, 195, 273-287.
- TANABE, K., MITA, T., JOMBART, T., ERIKSSON, A., HORIBE, S., PALACPAC, N., RANFORD-CARTWRIGHT, L., SAWAI, H., SAKIHAMA, N., OHMAE, H., NAKAMURA, M., FERREIRA, M. U., ESCALANTE, A. A., PRUGNOLLE, F., BJORKMAN, A., FARNERT, A., KANEKO, A., HORII, T., MANICA, A., KISHINO, H. & BALLOUX, F. 2010. Plasmodium falciparum accompanied the human expansion out of Africa. *Curr Biol*, 20, 1283-9.
- TANABE, K., MITA, T., PALACPAC, N. M., ARISUE, N., TOUGAN, T., KAWAI, S., JOMBART, T., KOBAYASHI, F. & HORII, T. 2013. Within-population genetic diversity of Plasmodium falciparum vaccine candidate antigens reveals geographic distance from a Central sub-Saharan African origin. *Vaccine*, 31, 1334-9.
- TANABE, K., MITA, T., RANFORD-CARTWRIGHT, L., PALACPAC, N., SAWAI, H., SAKIHAMA, N., HORIBE, S., OHMAE, H., NAKAMURA, M., FERREIRA, M. U., ESCALANTE, A., BJORKMAN, A., FARNERT, A., KANEKO, A., HORII, T., KISHINO, H. 2009. Geographical distribution of diversity in antigen genes and housekeeping genes of Plasmodium falciparum. *Unpublished*.
- TANABE, K., SAKIHAMA, N. & KANEKO, A. 2004. Stable SNPs in malaria antigen genes in isolated populations. *Science*, 303, 493.
- TANABE, K., SAKIHAMA, N., ROTH, I., BJORKMAN, A. & FARNERT, A. 2007a. High frequency of recombination-driven allelic diversity and temporal variation of Plasmodium falciparum msp1 in Tanzania. *Am J Trop Med Hyg*, 76, 1037-45.
- TANABE, K., SAKIHAMA, N., WALLIKER, D., BABIKER, H., ABDEL-MUHSIN, A. M., BAKOTE'E, B., OHMAE, H., ARISUE, N., HORII, T., ROTH, I., FARNERT, A., BJORKMAN, A. & RANFORD-

- CARTWRIGHT, L. 2007b. Allelic dimorphism-associated restriction of recombination in *Plasmodium falciparum* msp1. *Gene*, 397, 153-60.
- TANABE, K., ZOLLNER, G. E., SATTABONGKOT, J., KHUNTIRAT, B., HONMA, H., MITA, T., TSUBOI, T. AND COLEMAN, R. 2013. Genetic diversity of *Plasmodium falciparum* in an isolated village in western Thailand. *Unpublished*.
- WEBER, J. L., LEININGER, W. M. & LYON, J. A. 1986. Variation in the gene encoding a major merozoite surface antigen of the human malaria parasite *Plasmodium falciparum*. *Nucleic Acids Res*, 14, 3311-23.
- YAVO, W., KONATE, A., MAWILI-MBOUMBA, D. P., KASSI, F. K., TSHIBOLA MBUYI, M. L., ANGORA, E. K., MENAN, E. I. & BOUYOU-AKOTET, M. K. 2016. Genetic Polymorphism of msp1 and msp2 in *Plasmodium falciparum* Isolates from Cote d'Ivoire versus Gabon. *J Parasitol Res*, 2016, 3074803.
- YUAN, L., ZHAO, H., WU, L., LI, X., PARKER, D., XU, S., ZHAO, Y., FENG, G., WANG, Y., YAN, G., FAN, Q., YANG, Z. & CUI, L. 2013. *Plasmodium falciparum* populations from northeastern Myanmar display high levels of genetic diversity at multiple antigenic loci. *Acta Trop*, 125, 53-9.